

Tight Space-Approximation Tradeoff for the Multi-Pass Streaming Set Cover Problem

Sepehr Assadi*

University of Pennsylvania

sassadi@cis.upenn.edu

Abstract

We study the classic set cover problem in the streaming model: the sets that comprise the instance are revealed one by one in a stream and the goal is to solve the problem by making one or few passes over the stream while maintaining a sublinear space $o(mn)$ in the input size; here m denotes the number of the sets and n is the universe size. Notice that in this model, we are mainly concerned with the space requirement of the algorithms and hence do not restrict their computation time.

Our main result is a resolution of the space-approximation tradeoff for the streaming set cover problem: we show that any α -approximation algorithm for the set cover problem requires $\tilde{\Omega}(mn^{1/\alpha})$ space, even if it is allowed $\text{polylog}(n)$ passes over the stream, and even if the sets are arriving in a random order in the stream. This space-approximation tradeoff matches the best known bounds achieved by the recent algorithm of Har-Peled et al. (PODS 2016) that requires only $O(\alpha)$ passes over the stream in an adversarial order, hence settling the space complexity of approximating the set cover problem in data streams in a quite robust manner. Additionally, our approach yields tight lower bounds for the space complexity of $(1 - \varepsilon)$ -approximating the streaming maximum coverage problem studied in several recent works.

*Supported in part by National Science Foundation grants CCF-1552909, CCF-1617851, and IIS-1447470.

1 Introduction

The *set cover* problem is one of the most fundamental optimization problems in computer science, with a wide range of applications in various domains including data mining and information retrieval [2, 47], web host analysis [20], operation research [30], and many others. In this problem, we are given a collection of m sets from a universe $[n]$ and the goal is to output a smallest number of sets whose union is $[n]$, or in other words, *cover* the universe. The set cover problem is one of Karp’s original 21 NP-hard problems [38]. A simple greedy algorithm that iteratively picks the set that covers the most number of uncovered elements achieves a $(\ln n)$ -approximation [36, 48] and this is best possible unless $P = NP$ [24, 28, 41, 43].

The aforementioned results focus on the tradeoff between approximation guarantee and time complexity of the set cover problem. Nevertheless, in many settings, *space complexity* of the algorithms is crucial to optimize. A canonical example is in applications in big data analysis: in such settings, one would like to design algorithms capable of processing massive datasets using only few passes over the input and limited space. The well-established *streaming model* of computation [1, 44] precisely captures this setting.

In the *streaming set cover* problem, originally introduced by Saha and Getoor [47], the input sets are provided one by one in a stream and the algorithms are allowed to make a small number of passes over the stream while maintaining a sublinear space $o(mn)$ for processing the stream. The streaming set cover problem and the closely related maximum coverage problem have received quite a lot of attention in recent years [3, 5, 9, 18, 21, 23, 26, 32, 34, 42, 47]; we refer the reader to [3, 42] for a comprehensive summary of these results.

Particularly relevant to our work, Demaine et al. [23], have shown an α -approximation algorithm that uses $O(\alpha)$ passes over the stream and needs $\tilde{O}(mn^{\Theta(1/\log \alpha)})$ space. Recently, Har-Peled et al. [32] provide a significant improvement over this algorithm: they developed an α -approximation, $O(\alpha)$ -pass streaming algorithm that requires $\tilde{O}(mn^{\Theta(1/\alpha)})$ space. They further conjectured that the tradeoff between the number of passes and the space in their algorithm is almost tight: this is supported by a lower bound of $\tilde{\Omega}(mn^{1/2p})$ space for p -pass streaming algorithms that compute an *exact* set cover solution [32].

Notice however that the algorithm of [32] (and [23]) exhibits a somewhat unusual behavior: allowing a larger number of passes over the stream results in a weaker approximation guarantee obtained by the algorithm. This highlights the following natural question: can we achieve a (fixed) *constant* approximation in p -passes and $\tilde{O}(mn^{\Theta(1/p)})$ space? (a recent algorithm of Bateni et al. [9] achieves a fixed $\log n$ -approximation within these bounds.) In general, what is the *space-approximation tradeoff* for the streaming set cover problem if we consider algorithms that are allowed a relatively small number of passes, say up to $\text{polylog}(n)$, over the stream? This is precisely the question addressed in this work.

1.1 Our Contributions

Our main result is a tight resolution of the space-approximation tradeoff for the streaming set cover problem:

Result 1 (Main result, formalized as Theorem 1). *Any streaming α -approximation $\text{polylog}(n)$ -pass algorithm for the set cover problem requires $\tilde{\Omega}(mn^{1/\alpha})$ space even on random arrival streams. This lower bound applies even for the weaker goal of estimating the optimal value of the set cover instance (as opposed to finding the actual sets that cover the universe).*

Prior to our work, the best known lower bounds for *randomized multi-pass* streaming algorithms ruled out the possibility of $(\log n/2)$ -approximation in p passes and $o(m/p)$ space [45], and exact solution in p passes and $o(mn^{1/2p})$ space [32] (the latter holds only if $m = O(n)$). These results left open the possibility of obtaining, say, a 2-approximation in two passes or even an exact answer in $O(\log n)$ passes and $\tilde{O}(m)$ space. On the other hand, Result 1 smoothly extends the bounds in [45] to the whole range of approximation factors $\alpha = o(\log n)$, proving the first *super-linear* in m lower bound for approximating set cover in *multi-pass* streams. It also significantly improves the bounds in [32] to $\tilde{\Omega}(mn/p)$ (and all range of $m = \text{poly}(n)$) for p pass streaming algorithms that recover an exact answer¹.

As mentioned earlier, Har-Peled et al. [32] designed an α -approximation algorithm for the set cover problem that requires $\tilde{O}(mn^{\Theta(1/\alpha)})$ space (for some unspecified constant larger than 2 in the Θ -notation in the exponent). We can show that with proper modifications, this algorithm in fact only requires $\tilde{O}(mn^{1/\alpha})$ space (see Theorem 2), hence proving a *tight* upper bound for Result 1 (up to logarithmic factors). These results together resolve the space-approximation tradeoff for streaming set cover problem in multi-pass streams. It is worth mentioning that the space-approximation tradeoff for *single-pass* streaming algorithms of set cover has been previously resolved in [3].

Finally, we point out that the lower bound in Result 1 is quite *robust* in the sense that it holds even when the sets are arriving in a random order. This is particularly relevant to the streaming set cover problem as most known techniques for this problem are based on element and set sampling and a-priori one may expect that random arrival streams can facilitate the use of such techniques, resulting in better bounds than the ones achievable in adversarial streams. We point that in general, many streaming problems are known to be distinctly easier in random arrival streams compared to adversarial streams (see, e.g., [31, 37, 39]).

We further show an application of our techniques in establishing Result 1 to the *streaming maximum coverage* problem that has been studied in several recent works [4, 5, 9, 19, 27, 42, 47]. In this problem, we are given a collection of m sets from a universe $[n]$ and an integer $k \geq 1$, and the goal is to find k sets that cover the most number of elements in $[n]$. We prove that,

Result 2 (Formalized as Theorem 4). *Any streaming $(1 - \varepsilon)$ -approximation $\text{polylog}(n)$ -pass algorithm for the maximum coverage problem requires $\tilde{\Omega}(m/\varepsilon^2)$ space even on random arrival streams. This lower bound applies even for the case $k = O(1)$.*

Single-pass $(1 - \varepsilon)$ -approximation algorithms for this problem that use, respectively, $\tilde{O}(mk/\varepsilon^2)$ space and $\tilde{O}(m/\varepsilon^3)$ have been proposed recently in [9, 42], and [9]. Our Result 2 is hence *tight* for any $k = O(1)$ (up to logarithmic factors) and within an $O(1/\varepsilon)$ factor of the best upper bound for the larger values of k .

McGregor and Vu [42] have very recently proved an $\tilde{\Omega}(m)$ lower bound for $\text{polylog}(n)$ -pass streaming algorithms that approximate the maximum coverage problem to within a factor better than $(1 - 1/e)$ (a single-pass $(1 - 1/e)$ -approximation algorithm in $\tilde{O}(m)$ space is also developed in [9, 42]). The importance of Result 2 is thus in establishing the tight dependence on the parameter ε for this problem. This is important as $(1 - \varepsilon)$ -approximation algorithms for this problem for very small values of ε , i.e., $\varepsilon = 1/n^{\Omega(1)}$, are typically used as a sub-routine in approximating the streaming set cover problem in multiple passes [9, 23, 32] (see Section 3.4 for more details).

En route, we also obtain the following result which may be of independent interest: the communication complexity of computing an exact solution to the set cover problem or the maximum

¹Note that this result also implies that the “right” tradeoff between space and number of passes for obtaining an *exact* solution to the streaming set cover is in fact linear as opposed to exponential, i.e., n/p as opposed to $n^{1/p}$, as was previously shown in [32].

coverage problem in the two-player communication model is $\tilde{\Omega}(mn)$ bits (see Theorems 3 and 5). This improves upon the previous $\Omega(m)$ lower bounds of Nisan [45] (for set cover) and McGregor and Vu [42] (for maximum coverage). The two-player communication model for set cover has also been studied in [3, 18, 23, 32].

We conclude this section by highlighting the following important aspect of our lower bounds.

Remark 1.1. *In the hard instances we consider in proving Results 1 and 2, the minimum set cover size and the parameter k in maximum coverage are small constants and hence these instances admit a trivial poly-time algorithm in the classical (offline) setting. Our results hence establish the “hardness” of these instances under the space restrictions of the streaming model, independent of the NP-hardness of approximating these problems.*

1.2 Technical Overview

We focus here on providing a technical overview of the proof of Result 1 - Result 2 is also proven along similar lines. The starting point of our work is [3], which proved a tight space lower bound for single-pass streaming algorithms of set cover by analyzing the *one-way communication complexity* of this problem (see Section 2 for details on communication complexity).

The overall approach of [3] can be summarized as follows. Consider a communication problem whereby Alice is given a collection of sets S_1, \dots, S_m , Bob is given a set T , and they need to compute an α -approximation of the set cover instance (S_1, \dots, S_m, T) in the one-way communication model. The input to the players are correlated in that there exists a set S_{i^*} in Alice’s collection which together with Bob’s set T cover the whole universe except for a single element. However, if the content of the set S_{i^*} is unknown to Bob, i.e., Alice’s message does not reveal almost all S_{i^*} , Bob needs to cover $[n] \setminus T$ (which is a subset of S_{i^*} except for one element) with sets other than S_{i^*} to ensure that the single element outside S_{i^*} is covered. The collection S_1, \dots, S_m is designed to satisfy the so-called *r -covering* property [41] that states that no small collection of S_i ’s set can cover another set S_j entirely², hence forcing Bob to use many sets to cover the universe. The authors then use the *information complexity* paradigm to reduce the set cover problem on this distribution to multiple instances of a simpler problem (called the *Trap* problem) and prove a lower bound for this new problem.

In this paper, we extend this approach to lower bound the two-way communication complexity of the set cover problem and ultimately obtain the desired lower bound in Result 1 for multi-pass streaming algorithms. To do this, we need to address the following issues:

First, the type of distribution used in [3] is clearly not suitable for proving lower bounds in the two-way model. In particular, we need a distribution with both Alice and Bob having $\Omega(m)$ sets and additionally, no clear “signal” to either party as which of the sets are more important, i.e., correspond to the sets S_{i^*} and T in the above distribution. To achieve this, we employ the *r -covering* property in a novel way: we first design a collection of sets Z_1, \dots, Z_m such that no collection of α sets Z_i ’s can cover the universe $[n]$ unless they contain a single set Z_{i^*} which is in fact equal to $[n]$ already (for remaining sets Z_i , we have $|Z_i| \approx n - n^{1-1/\alpha}$). Next, we decompose each Z_i into two sets S_i and T_i and provide Alice with S_i , and Bob with T_i . This way, the sets S_{i^*} and T_{i^*} form a set cover of size two, and the *r -covering* property ensures that no other collection of α pairs (S_i, T_i) can cover the universe; we further prove that “mix and matching” the sets (i.e., picking S_i but not T_i or vice versa) in the solution is not helpful either, hence implying that any α -approximation algorithm for set cover needs to find the sets S_{i^*} and T_{i^*} .

²It is worth mentioning that essentially all known lower bounds for the streaming set cover problem, on their core, are based on some variant of this *r -covering* property; see [18] for more details.

The next step is to prove the lower bound for the above distribution. Unlike the lower bound in the one-way model that was based on hiding the content of the set S_{i^*} , here we need to argue that in fact the index i^* itself is hidden from the players (as otherwise, one more round of communication can reveal the content of the sets S_{i^*} and T_{i^*} as well). Similar to [3], we also use the information complexity paradigm to prove the communication lower bound for this distribution. We embed different instances of the well-known *set disjointness* problem in each pair (S_i, T_i) such that all embedded instances are *intersecting* except for the instance for S_{i^*} and T_{i^*} which is *disjoint*. As we seek a direct-sum style argument for two-way protocols, we need a more careful argument than the one in [3] that was tailored for one-way protocols. In particular, we now use the notion of *internal* information complexity (as opposed to *external* information complexity used in [3]) that allows us to use the powerful techniques developed in [8, 10, 13] to obtain the direct-sum result.

Finally, we need to lower bound the information complexity of the set disjointness problem on the specific distribution induced by the set cover instances. The set cover distribution is designed in a way to ensure that the distribution of underlying set disjointness instances matches the known hard input distributions for this problem. However, there is a subtlety here; known information complexity lower bounds for set disjointness (that we are aware of) are all over distributions that are supported only on disjoint sets, i.e., **Yes**-instances of the problem (see, e.g., [6, 11, 49])³. However, for our purpose, we need to lower bound the information cost of set disjointness protocols on distributions that are intersecting. We achieve this using an application of the “information odometer” of [14] (and subsequent work in [29]) to relate the information cost of the protocols on **Yes** and **No** instances of the problem together and obtain the result.

We are not done though, as we seek a lower bound for random arrival streams and for this, we extend the previous communication complexity lower bound to the case when the input sets are partitioned randomly across the players, in a similar way as done in previous work [3] (itself based on [15]). There are however some technical differences needed to execute this approach in our two-way communication model in compare to the one-way model in [3] (see Lemma 3.7 for details).

2 Preliminaries

Notation. For any integer $a \geq 1$, we let $[a] := \{1, \dots, a\}$. We say that a set $S \subseteq [n]$ with $|S| = s$ is a *s-subset* of $[n]$. For a k -dimensional tuple $X = (X_1, \dots, X_k)$ and index $i \in [k]$, we define $X^{<i} := (X_1, \dots, X_{i-1})$ and $X^{-i} := (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_k)$.

We use capital letters to denote random variables. For a random variable A , $\text{SUPP}(A)$ denotes the support of A and $|A| := \log |\text{SUPP}(A)|$. We use “ $A \perp B \mid C$ ” to mean that the random variables A and B are independent conditioned on C . The notation “ $A \in_R U$ ” indicates that A is chosen uniformly at random from the set U .

We denote the *Shannon Entropy* of a random variable A by $\mathbb{H}(A)$ and the *mutual information* of two random variables A and B by $\mathbb{I}(A : B) = \mathbb{H}(A) - \mathbb{H}(A \mid B) = \mathbb{H}(B) - \mathbb{H}(B \mid A)$. If the distribution \mathcal{D} of the random variables is not clear from the context, we use $\mathbb{H}_{\mathcal{D}}(A)$ (resp. $\mathbb{I}_{\mathcal{D}}(A : B)$). Appendix A summarizes the relevant information theory tools that we use in this paper.

Concentration bounds. We use the following standard version of Chernoff bound (see, e.g., [25]).

³We remark that this is not just a coincidence and in fact is crucial for performing the typical reduction to the AND problem used in proving the lower bound for set disjointness, see, e.g., [49] for more details.

Proposition 2.1. Let X_1, \dots, X_n be n independent random variables taking values in $[0, 1]$ and let $X := \sum_{i=1}^n X_i$. Then, for any $0 \leq \varepsilon \leq 1$,

$$\mathbb{P}(|X - \mathbb{E}[X]| > \varepsilon \cdot \mathbb{E}[X]) \leq 2 \cdot \exp\left(-\frac{\varepsilon^2 \cdot \mathbb{E}[X]}{2}\right)$$

We also prove the following useful auxiliary lemma that upper bounds the number of elements that a collection of large random sets can cover.

Lemma 2.2. Let $\mathcal{S} = \{S_1, \dots, S_k\}$ be a collection of $(n - s)$ -subsets of $[n]$ that are chosen independently and uniformly at random. Suppose $U \subseteq [n]$ is another set chosen independent of \mathcal{S} ; if $k = o(e^s)$, then,

$$\mathbb{P}\left(|U \setminus (S_1 \cup \dots \cup S_k)| < \frac{|U|}{2} \cdot \left(\frac{s}{2n}\right)^k\right) < 2 \cdot \exp\left(-\frac{|U|}{8} \cdot \left(\frac{s}{2n}\right)^k\right)$$

We first briefly explain the bounds in Lemma 2.2. Note that each element $e \in [n]$, is not covered by a set $S_i \in \mathcal{S}$ w.p. $\frac{s}{n}$ (as S_i is a random set of size $(n - s)$). Moreover, since the sets are chosen independent of each other, the probability that e is not covered by \mathcal{S} is $\left(\frac{s}{n}\right)^k$. Hence, in expectation $|U| \cdot \left(\frac{s}{n}\right)^k$ elements in U are not covered by \mathcal{S} . We then wish to argue, by means of some concentration bound, that with a very high probability the number of elements not covered by \mathcal{S} is at least half of this number (notice that the bounds in the lemma statement are quite similar but not exactly equal to this quantity).

However, there is an important subtlety here. The random variables defined in the above process are *negatively correlated* and hence one cannot readily use a Chernoff-Hoeffding bound (or even similar variants defined for negatively correlated random variables) to bound this probability. This is because we need to bound the probability of the sum of these random variable being too small as opposed to being too large which already follows from known results (see, e.g., [33, 46])⁴. In the following, we show how to get around this using a careful coupling argument.

Proof of Lemma 2.2. For any element $e \in U$, define the random variable $X_e \in \{0, 1\}$ which is 1 iff $e \notin S_1 \cup \dots \cup S_k$. Define $X := \sum_{e \in U} X_e$; notice that X denotes the number of elements in U that are not covered by \mathcal{S} . Our goal is then to lower bound the value of X . Note that the random variables X_e are *negatively correlated* and hence, as stated earlier, we *cannot* use Chernoff bound (or its generalizations to negatively correlated random variables) to *lower bound* the value of X .

To get around this, we slightly change the distribution each set is chosen from, prove the result in that case, and then relate that distribution to the original distribution of the sets in \mathcal{S} . Formally, let \mathcal{D} be the distribution of from which the sets in \mathcal{S} are chosen. Consider the following distribution \mathcal{D}' : we create each set S_i (for $i \in [k]$) by removing each element in $[n]$ from S_i independently and uniformly at random w.p. $p = \frac{s}{2n}$.

We lower bound the value of the random variable X under this new distribution. We first have,

$$\mathbb{E}_{\mathcal{D}'}[X] = \sum_{e \in U} \mathbb{P}_{\mathcal{D}'}(X_e = 1) = |U| \cdot p^k = |U| \cdot \left(\frac{s}{2n}\right)^k$$

For simplicity, define $\eta := |U| \cdot \left(\frac{s}{2n}\right)^k$. An important property of \mathcal{D}' is that now all random variables X_e are *independent* of each other. Hence, we can apply Chernoff bound as follows,

$$\mathbb{P}_{\mathcal{D}'}(X < \eta/2) = \mathbb{P}_{\mathcal{D}'}(X < \mathbb{E}[X]/2) \leq e^{-\eta/8} \quad (1)$$

⁴Note that in general, Chernoff bound type inequalities do not hold for bounding the sum of negatively random variables from below.

We now argue that $\mathbb{P}_{\mathcal{D}}(X < \eta/2)$ is in fact very close to $\mathbb{P}_{\mathcal{D}'}(X < \eta/2)$.

Fix a set $S_i \in \mathcal{S}$. For each $e \in [n]$, define a random variable $Y_e \in \{0, 1\}$ which is 1 iff $e \notin S_i$. Let $Y = \sum_{e \in [n]} Y_e$, i.e., the number of elements missing from S_i . Note that $\mathbb{E}_{\mathcal{D}'}[Y] = \frac{s}{2}$. Under the distribution \mathcal{D}' , for each set $S_i \in \mathcal{S}$, each element $e \in [n]$ belongs to S_i independently; hence a simple application of Chernoff bound ensures that:

$$\mathbb{P}_{\mathcal{D}'}(|S_i| < (n - s)) = \mathbb{P}_{\mathcal{D}'}(Y > 2 \cdot \mathbb{E}[Y]) < e^{-s/2} \quad (2)$$

Define \mathcal{E} as the event that all sets S_i has size at least $(n - s)$; by Eq (2) and a union bound, $\mathbb{P}_{\mathcal{D}'}(\mathcal{E}) \geq 1 - k \cdot e^{-s/2} \geq \frac{1}{2}$ (as $k = o(e^s)$). Notice that to sample a set system from \mathcal{D} , we can first sample a set system from $\mathcal{D}' \mid \mathcal{E}$ and then make the size of each set exactly equal to $(n - s)$ by removing the extra elements uniformly at random; this process does not increase the coverage of the original set system sampled from \mathcal{D}' (or equivalently decrease the value of X). Hence,

$$\mathbb{P}_{\mathcal{D}}(X < \eta/2) \leq \mathbb{P}_{\mathcal{D}'}(X < \eta/2 \mid \mathcal{E}) \leq \frac{\mathbb{P}_{\mathcal{D}'}(X < \eta/2)}{\mathbb{P}_{\mathcal{D}'}(\mathcal{E})} \leq 2 \cdot e^{-\eta/8}$$

By substituting the value of η , we obtain the desired bound. \blacksquare

2.1 Communication Complexity and Information Complexity

Communication complexity and information complexity play an important role in our lower bound proofs. We now provide necessary definitions for completeness.

Communication complexity. Our lower bounds for streaming algorithms are established via communication complexity lower bounds. We use standard definitions of the *two-party communication* model introduced by Yao [50]; see [40] for an extensive overview of communication complexity.

Let P be a relation with domain $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. Alice receives an input $X \in \mathcal{X}$ and Bob receives $Y \in \mathcal{Y}$, where (X, Y) are chosen from a joint distribution \mathcal{D} over $\mathcal{X} \times \mathcal{Y}$. They communicate with each other by exchanging messages such that each message depends only on the private input of the player sending the message and the already communicated messages. The last message communicated is the answer Z such that $(X, Y, Z) \in P$. We allow players to have access to both public and private randomness.

We use π to denote a protocol used by the players. We always assume that the protocol π can be randomized (using both public and private randomness), *even against a prior distribution \mathcal{D} of inputs*. For any $0 < \delta < 1$, we say π is a δ -error protocol for P over a distribution \mathcal{D} , if the probability that for an input (X, Y) , π outputs some Z where $(X, Y, Z) \notin P$ is at most δ (the probability is taken over the randomness of both the distribution and the protocol).

Definition 1. The communication cost of a protocol π for a problem P on an input distribution \mathcal{D} , denoted by $\|\pi\|$, is the worst-case bit-length of the transcript communicated between Alice and Bob in the protocol π , when the inputs are chosen from \mathcal{D} .

The communication complexity $\text{CC}_{\mathcal{D}}^{\delta}(P)$ of a problem P with respect to a distribution \mathcal{D} is the minimum communication cost of a δ -error protocol π over \mathcal{D} .

Information complexity. There are several possible definitions of information complexity of a communication problem that have been considered depending on the application (see, e.g., [6–8, 13, 17]). We use the notion of *internal information complexity* [8] that measures the average amount of (Shannon) information each player learns about the input of the other player by observing the transcript of the protocol. Formally,

Definition 2. Consider an input distribution \mathcal{D} and a protocol π (for some problem P). Let $(X, Y) \sim \mathcal{D}$ be the input of Alice and Bob and assume $\Pi := \Pi(X, Y)$ denotes the transcript of the protocol concatenated with the public randomness R used by π . The (internal) information cost $\text{ICost}_{\mathcal{D}}(\pi)$ of a protocol π with respect to \mathcal{D} is then $\mathbb{I}_{\mathcal{D}}(\Pi : X | Y) + \mathbb{I}_{\mathcal{D}}(\Pi : Y | X)$.

The information complexity $\text{IC}_{\mathcal{D}}^{\delta}(P)$ of P with respect to a distribution \mathcal{D} is the minimum $\text{ICost}_{\mathcal{D}}(\pi)$ taken over all δ -error protocols π for P over \mathcal{D} .

Note that any public coin protocol is a distribution over private coins protocols, obtained by first using public randomness to sample a random string $R = r$ and then running the corresponding private coin protocol π^r . We also use Π^r to denote the transcript of the protocol π^r . We have the following well-known claim.

Claim 2.3. For any distribution \mathcal{D} and any protocol π , let R be the public randomness used in π ; then, $\text{ICost}_{\mathcal{D}}(\pi) = \mathbb{I}_{\mathcal{D}}(\Pi : X | Y, R) + \mathbb{I}_{\mathcal{D}}(\Pi : Y | X, R)$.

Proof.

$$\begin{aligned}
\text{ICost}_{\mathcal{D}}(\pi) &= \mathbb{I}(\Pi : X | Y) + \mathbb{I}(\Pi : Y | X) \\
&= \mathbb{I}(\Pi, R : X | Y) + \mathbb{I}(\Pi, R : Y | X) \\
&\quad (\Pi \text{ denotes the transcript and the public randomness}) \\
&= \mathbb{I}(R : X | Y) + \mathbb{I}(\Pi : X | Y, R) + \mathbb{I}(R : Y | X) + \mathbb{I}(\Pi : Y | X, R) \\
&\quad (\text{chain rule of mutual information, Fact A.1-(4)}) \\
&= \mathbb{I}(\Pi : X | Y, R) + \mathbb{I}(\Pi : Y | X, R)
\end{aligned}$$

The last equality is because $\mathbb{I}(R : X | Y) = \mathbb{I}(R : Y | X) = 0$ since $R \perp X, Y$ and Fact A.1-(2). ■

The following well-known proposition relates communication complexity and internal information complexity (see, e.g., [13] for a proof).

Proposition 2.4. For any distribution \mathcal{D} and any protocol π : $\text{ICost}_{\mathcal{D}}(\pi) \leq \|\pi\|$. Moreover, for any parameter $0 < \delta < 1$: $\text{IC}_{\mathcal{D}}^{\delta}(P) \leq \text{CC}_{\mathcal{D}}^{\delta}(P)$.

2.2 The Set Disjointness Problem

We shall use the well-known *set-disjointness* communication problem (denoted by Disj) in proving Result 1. Fix an integer $t \geq 1$; in Disj_t , Alice and Bob are given two sets $A \subseteq [t]$ and $B \subseteq [t]$, and their goal is to return **Yes** if $A \cap B = \emptyset$ and **No** otherwise.

The following is a known hard distribution for Disj_t .

Distribution $\mathcal{D}_{\text{Disj}}$. A hard input distribution for Disj_t .

- Start with $A = B = [t]$.
- For each element $e \in [t]$ independently: w.p. 1/3 drop e from both A and B , w.p. 1/3 drop e from A , and w.p. 1/3 drop e from B .
- Pick $Z \in_R \{0, 1\}$ uniformly at random. If $Z = 1$, pick a uniformly at random element $e^* \in [t]$ and let A and B both contain e^* (if $Z = 0$, keep the sets as before).

We further use $\mathcal{D}_{\text{Disj}}^Y$ and $\mathcal{D}_{\text{Disj}}^N$ to denote, respectively, the distribution of **Yes** and **No** instances of Disj on $\mathcal{D}_{\text{Disj}}$; in other words, $\mathcal{D}_{\text{Disj}}^Y := (\mathcal{D}_{\text{Disj}} | Z = 0)$ and $\mathcal{D}_{\text{Disj}}^N := (\mathcal{D}_{\text{Disj}} | Z = 1)$.

The following proposition on the information complexity of Disj is well-known (see, e.g., [6, 11]).

Proposition 2.5. *For any $\delta < 1/2$ and any δ -error protocol π_{Disj} of Disj_t on the distribution $\mathcal{D}_{\text{Disj}}$,*

$$\text{ICost}_{\mathcal{D}_{\text{Disj}}}^Y(\pi_{\text{Disj}}) = \Omega(t).$$

3 The Space-Approximation Tradeoff for Set Cover

We prove our main result on the space-approximation tradeoff for the streaming set cover problem in this section. Formally,

Theorem 1. *For any $\alpha = o(\log n / \log \log n)$, $m = \text{poly}(n)$, and $p \geq 1$, any randomized algorithm that can make p passes over any collection of m subsets of $[n]$ presented in a random order stream and outputs an α -approximation to the optimal value of the set cover problem w.p. larger than $3/4$ (over the randomness of both the stream order and the algorithm) must use $\tilde{\Omega}(mn^{1/\alpha}/p)$ space.*

Theorem 1 formalizes Result 1 in the introduction. We further prove that the tradeoff achieved in Theorem 1 is in fact tight up to logarithmic factors; this is achieved by performing some proper modifications to the algorithm of [32]. Formally,

Theorem 2. *There exists a streaming algorithm that for any integer $\alpha \geq 1$, and any parameter $\varepsilon > 0$, with high probability, computes an $(\alpha + \varepsilon)$ -approximation to the streaming set cover problem using $(2\alpha + 1)$ passes over the stream in adversarial order and $\tilde{O}(mn^{1/\alpha}/\varepsilon^2 + n/\varepsilon)$ space.*

We emphasize that the main contribution of the paper is in proving Theorem 1; we mainly present Theorem 2 to prove a matching upper bound on the bounds in Theorem 1, hence establishing a tight space-approximation tradeoff for the streaming set cover problem.

The rest of this section is mainly devoted to the proof of Theorem 1. We start by introducing some notation. In Section 3.1, we introduce a hard input distribution for the set cover problem in adversarial streams. We prove a lower bound for this distribution in Section 3.2. We extend this lower bound to random arrival streams in Section 3.3 and finish the proof of Theorem 1. Section 3.4 contains the proof of Theorem 2.

Notation. To prove Theorem 1, we prove a lower bound on the communication complexity of the set cover problem: Fix a (sufficiently large) value for n , $m = \text{poly}(n)$, and $\alpha = o(\log n / \log \log n)$; in this section, **SetCover** refers to the problem of α -approximating the optimal value of the set cover problem with $2m$ sets⁵ defined over the universe $[n]$ in the two-player communication model, whereby the sets are partitioned between Alice and Bob.

3.1 A Hard Input Distribution for SetCover

Let t be an integer to be determined later; we use the distribution $\mathcal{D}_{\text{Disj}}$ for Disj_t (introduced in Section 2.2) to design a hard input distribution for **SetCover**. Before that, we need a simple definition.

Definition 3 (Mapping-extension). *For the two sets $[t]$ and $[n]$, we define a mapping-extension of $[t]$ to $[n]$ as a function $f : [t] \mapsto 2^{[n]}$, whereby for each $i \in [t]$, $f(i) \subseteq [n]$ is mapped to n/t unique elements in $[n]$. Similarly, for any set $A \subseteq [t]$, we abuse the notation and define $f(A) := \bigcup_{i \in A} f(i)$.*

⁵To simplify the exposition, we use $2m$ instead of m as the number of sets.

We are now ready to define our hard input distribution for **SetCover**.

Distribution \mathcal{D}_{SC} . A hard input distribution for **SetCover**.

Notation. Let $t := 2^{-15} \cdot \left(\frac{n}{\log m}\right)^{\frac{1}{\alpha}}$ and \mathcal{F} be the set of all mapping-extensions of $[t]$ to $[n]$.

- For each $i \in [m]$:
 - Let $(A_i, B_i) \sim \mathcal{D}_{\text{Disj}}^N$ for Disj_t and pick $f_i \in_R \mathcal{F}$ uniformly at random.
 - Let $S_i = [n] \setminus f_i(A_i)$ and $T_i := [n] \setminus f_i(B_i)$.
- Pick $\theta \in_R \{0, 1\}$ uniformly at random. If $\theta = 0$, do nothing, otherwise:
 - Sample $i^* \in_R [m]$ uniformly at random.
 - Resample $(A_{i^*}, B_{i^*}) \sim \mathcal{D}_{\text{Disj}}^Y$ for Disj_t and redefine S_{i^*} and T_{i^*} as before using the new pair (A_{i^*}, B_{i^*}) .
- Let the input to Alice and Bob be $\mathcal{S} := \{S_i\}_{i \in [m]}$ and $\mathcal{T} := \{T_i\}_{i \in [m]}$, respectively.

In the following, we use Z to denote any set in $\mathcal{S} \cup \mathcal{T}$, i.e., when it is not relevant whether it belongs to \mathcal{S} or \mathcal{T} . For a collection of sets $\mathcal{Z} = \{Z_1, \dots, Z_\ell\}$, we use $C(\mathcal{Z})$ to denote the set of elements that \mathcal{Z} covers, i.e., $C(\mathcal{Z}) := \bigcup_{i=1}^\ell Z_i$. We say that \mathcal{Z} is a *singleton-collection*, if for any $i \in [m]$, at least one of S_i or T_i is *not* present in \mathcal{Z} . In contrast, we say that \mathcal{Z} is a *pair-collection*, if for all $i \in [m]$, $S_i \in \mathcal{Z}$ iff $T_i \in \mathcal{Z}$ as well.

Remark 3.1. A few remarks are in order:

- (i) *W.h.p., for any $i \in [m]$, $|S_i| = 2n/3 \pm o(n)$ and $|T_i| = 2n/3 \pm o(n)$.
(Proof. follows from the definition of the distribution $\mathcal{D}_{\text{Disj}}$ and Chernoff bound).*
- (ii) *For any $i \in [m]$, conditioned on $|S_i| = \ell$, the set S_i is chosen uniformly at random from all ℓ -subsets of $[n]$; similarly for T_i*
- (iii) *For any $i \in [m]$, $S_i \cup T_i = [n] \setminus f_i(A_i \cap B_i)$. Moreover, whenever $(A_i, B_i) \sim \mathcal{D}_{\text{Disj}}^N$, the set $f_i(A_i \cap B_i)$ is a (n/t) -subset of $[n]$ chosen uniformly at random.
(Proof. the first part follows from the fact that f_i maps each $j \in [t]$ to unique elements; the second part is by the random choice of $f_i \in_R \mathcal{F}$ and the fact that $|A_i \cap B_i| = 1$ in this case).*
- (iv) *Whenever $\theta = 0$, for any $i \neq j$, the sets $Z_i \in \{S_i, T_i\}$ and $Z_j \in \{S_j, T_j\}$ are chosen independent of each other ($Z_i \perp Z_j$).*

Let $\text{opt}(\mathcal{S}, \mathcal{T})$ denote the size of an optimal set cover in the instance $(\mathcal{S}, \mathcal{T})$. It follows from Remark 3.1-(iii) that whenever $\theta = 1$ in the distribution \mathcal{D}_{SC} , $\text{opt}(\mathcal{S}, \mathcal{T}) = 2$; simply take S_{i^*} and T_{i^*} and since $A_{i^*} \cap B_{i^*} = \emptyset$, they cover the whole universe. In the following, we prove that when $\theta = 0$, $\text{opt}(\mathcal{S}, \mathcal{T})$ is relatively large. This implies that any α -approximation protocol for **SetCover** has to essentially determine the value of θ . In the next section, we prove that this task requires a large communication by the players.

Lemma 3.2. For $(\mathcal{S}, \mathcal{T}) \sim \mathcal{D}_{\text{SC}}$:

$$\mathbb{P}(\text{opt}(\mathcal{S}, \mathcal{T}) > 2\alpha \mid \theta = 0) = 1 - o(1).$$

Proof. Let \mathcal{C} be any collection of 2α sets from $(\mathcal{S}, \mathcal{T})$. We bound the probability that \mathcal{C} covers the universe $[n]$ entirely, i.e., is a feasible set cover, and then use a union bound on all possible choices for \mathcal{C} to finalize the proof. In the following, we condition on the event \mathcal{E}_1 that states that $|S_i| \leq 3n/4$ and $|T_i| \leq 3n/4$ for all $i \in [m]$ (which happens with probability $1 - o(1)$ by Remark 3.1-(i)).

Partition the collection \mathcal{C} into a pair-collection \mathcal{C}_P , and a singleton-collection \mathcal{C}_S (this partitioning is always possible and unique by definition). We first lower bound the number of elements that are not covered by the singleton-collection:

Claim 3.3. $\mathbb{P}\left(\left|\overline{C(\mathcal{C}_S)}\right| \leq \frac{n}{2^{6\alpha+1}} \mid \mathcal{E}_1\right) \leq 1 - \frac{1}{m^{\omega(\alpha)}}.$

Proof. Let $\mathcal{C}_S := \{Z_1, \dots, Z_k\}$; clearly $k = |\mathcal{C}_S| \leq |\mathcal{C}| = 2\alpha$. Without loss of generality, we assume that $k = 2\alpha$. By conditioning on the event \mathcal{E}_1 and Remark 3.1-(ii), we know that each Z_i is an ℓ_i -subset of $[n]$, for some $\ell_i \leq 3n/4$, chosen uniformly at random from all ℓ_i -subsets of $[n]$. Again without loss of generality, we simply increase the size of each Z_i so that they all have size exactly $3n/4$. Moreover, since no two sets S_i and T_i are both simultaneously present in \mathcal{C}_S , by Remark 3.1-(iv), all sets in \mathcal{C}_S are chosen independent of each other.

Consequently, by Lemma 2.2, for $U = [n]$, $s = n/4$, and collection \mathcal{C}_S , we have,

$$\mathbb{P}\left(\left|\overline{C(\mathcal{C}_S)}\right| < \frac{n}{2} \cdot \left(\frac{1}{8}\right)^{2\alpha} \mid \mathcal{E}_1\right) < 2 \cdot \exp\left(-\frac{n}{8} \cdot \left(\frac{1}{8}\right)^{2\alpha}\right)$$

A simplification of the above equation, plus using the fact that $\alpha = o(\log n / \log \log n)$, and hence $n/2^{\Theta(\alpha)} = \omega(\alpha \log m)$, proves the final result. \blacksquare

Let \mathcal{E}_2 be the event that $\left|\overline{C(\mathcal{C}_S)}\right| \geq \frac{n}{2^{6\alpha+1}}$; in the following, we condition on this event. Now consider the sets in the pair-collection \mathcal{C}_P . For any pair $(S_i, T_i) \in \mathcal{C}_P$, we define $C_i := S_i \cup T_i$. Note that there are at most α different possible sets C_i . By Remark 3.1-(iii), the sets C_i 's are random sets of size $(n - n/t)$, and by Remark 3.1-(iv), they are chosen independent of each other. By Lemma 2.2, for $U = \overline{C(\mathcal{C}_S)}$, $s = n/t$, and collection of sets C_i 's, we have,

$$\mathbb{P}(U \setminus (C(\mathcal{C}_P))) = \emptyset \mid \mathcal{E}_1, \mathcal{E}_2) \leq 2 \cdot \exp\left(-\frac{n}{2^{6\alpha+4}} \cdot \left(\frac{1}{2t}\right)^\alpha\right) \leq \frac{1}{m^{3\alpha}}$$

We can now conclude,

$$\begin{aligned} \mathbb{P}(\text{opt}(\mathcal{S}, \mathcal{T}) \leq 2\alpha) &\leq \mathbb{P}(\overline{\mathcal{E}_1}) + \mathbb{P}(\exists \mathcal{C} \text{ that covers } [n] \mid \mathcal{E}_1) \\ &\leq \mathbb{P}(\overline{\mathcal{E}_1}) + \sum_{\mathcal{C}} (\mathbb{P}(\overline{\mathcal{E}_2} \mid \mathcal{E}_1) + \mathbb{P}(C(\mathcal{C}) = [n] \mid \mathcal{E}_1, \mathcal{E}_2)) \\ &\leq o(1) + \binom{m}{2\alpha} \cdot \left(\frac{1}{m^{\omega(\alpha)}} + \frac{1}{m^{3\alpha}}\right) = o(1) \end{aligned}$$

proving the lemma. \blacksquare

3.2 The Lower Bound for the Distribution \mathcal{D}_{SC}

Throughout this section, fix π_{SC} as a δ -error protocol for SetCover on the distribution \mathcal{D}_{SC} . We first show that protocol π_{SC} is essentially solving m copies of the Disj_t problem on the distribution $\mathcal{D}_{\text{Disj}}$ (for the parameter t the distribution \mathcal{D}_{SC}) and then use a direct-sum style argument (similar in spirit to the ones in [8, 10, 13]) to argue that the information cost of π_{SC} shall be m times larger

than the information complexity of solving Disj_t . However, to make the direct-sum argument work, we can only consider π_{SC} on the distribution $\mathcal{D}_{\text{SC}} \mid \theta = 0$, i.e., when *all* underlying Disj_t instances are sampled from $\mathcal{D}_{\text{Disj}}^{\text{N}}$. Consequently, we can only lower bound the information cost of π_{SC} based on the information complexity of Disj_t on the distribution $\mathcal{D}_{\text{Disj}}^{\text{N}}$.

Lemma 3.4. *There exists a $(\delta + o(1))$ -protocol π_{Disj} for Disj_t on the distribution $\mathcal{D}_{\text{Disj}}$ such that:*

1. $\text{ICost}_{\mathcal{D}_{\text{Disj}}^{\text{N}}}(\pi_{\text{Disj}}) = \frac{O(1)}{m} \cdot \text{ICost}_{\mathcal{D}_{\text{SC}}}(\pi_{\text{SC}})$.
2. $\|\pi_{\text{Disj}}\| = \|\pi_{\text{SC}}\|$.

Proof. We design the protocol π_{Disj} as follows:

Protocol π_{Disj} . The protocol for solving Disj_t using a protocol π_{SC} for **SetCover**.

Input: An instance $(A, B) \sim \mathcal{D}_{\text{Disj}}$. **Output:** **Yes** if $A \cap B = \emptyset$ and **No** otherwise.

1. Using public randomness, the players sample an index $i^* \in_R [m]$ and m mapping-extensions f_1, \dots, f_m independently and uniformly at random from \mathcal{F} .
2. Using public randomness, the players sample the sets $A^{<i^*}$ and $B^{>i^*}$ each from $\mathcal{D}_{\text{Disj}}^{\text{N}}$ independently.
3. Using private randomness, Alice samples the sets $A^{>i^*}$ such that $(A_j, B_j) \sim \mathcal{D}_{\text{Disj}}^{\text{N}}$ (for all $j > i^*$); similarly Bob samples the sets $B^{<i^*}$.
4. The players construct the collections $\mathcal{S} := \{S_1, \dots, S_m\}$ and $\mathcal{T} := \{T_1, \dots, T_m\}$ by setting $S_i := [n] \setminus f_i(A_i)$ and $T_i := [n] \setminus f_i(B_i)$ (exactly as in distribution \mathcal{D}_{SC}).
5. The players solve the **SetCover** instance using π_{SC} and output **No** iff π_{SC} estimates $\text{opt}(\mathcal{S}, \mathcal{T}) \leq 2\alpha$ and **Yes** otherwise.

It is easy to see that the distribution of instances $(\mathcal{S}, \mathcal{T})$ created in the protocol π_{Disj} matches the distribution \mathcal{D}_{SC} for **SetCover** exactly. Moreover, by Lemma 3.2, $\text{opt}(\mathcal{S}, \mathcal{T}) > 2\alpha$ w.p. $1 - o(1)$, whenever $(A, B) \sim \mathcal{D}_{\text{Disj}}^{\text{N}}$ and $\text{opt}(\mathcal{S}, \mathcal{T}) = 2$ whenever $(A, B) \sim \mathcal{D}_{\text{Disj}}^{\text{Y}}$. Consequently, since π_{SC} is an α -approximation protocol,

$$\mathbb{P}_{\mathcal{D}_{\text{Disj}}}(\pi_{\text{Disj}} \text{ errs}) \leq \mathbb{P}_{\mathcal{D}_{\text{SC}}}(\pi_{\text{SC}} \text{ errs}) + o(1) \leq \delta + o(1)$$

and hence π_{Disj} is indeed a $(\delta + o(1))$ -error protocol for Disj on the distribution $\mathcal{D}_{\text{Disj}}$. Moreover, it is clear that the communication cost of π_{Disj} is at most the communication cost of π_{SC} . We now prove the bound on the information cost of this protocol.

Our goal is to bound the information cost of π_{Disj} whenever the instance (A, B) is sampled from $\mathcal{D}_{\text{Disj}}^{\text{N}}$. Let F be a random variable denoting the tuple (f_1, \dots, f_m) , I be a random variable for i^* and R be the set of public randomness used by the players. By Claim 2.3,

$$\text{ICost}_{\mathcal{D}_{\text{Disj}}^{\text{N}}}(\pi_{\text{Disj}}) = \mathbb{I}_{\mathcal{D}_{\text{Disj}}^{\text{N}}}(\Pi_{\text{Disj}} : A \mid B, R) + \mathbb{I}_{\mathcal{D}_{\text{Disj}}^{\text{N}}}(\Pi_{\text{Disj}} : B \mid A, R)$$

We now bound the first term in the RHS above (the second term can be bounded exactly the same).

$$\mathbb{I}_{\mathcal{D}_{\text{Disj}}^{\text{N}}}(\Pi_{\text{Disj}} : A \mid B, R) = \mathbb{I}_{\mathcal{D}_{\text{Disj}}^{\text{N}}}(\Pi_{\text{Disj}} : A \mid B, R, I) \quad (I \text{ is chosen using public randomness})$$

$$\begin{aligned}
&= \sum_{i=1}^m \mathbb{P}(I = i) \cdot \mathbb{I}_{\mathcal{D}_{\text{Disj}}^{\mathbf{N}}}(\Pi_{\text{Disj}} : A_i \mid B_i, A^{<i}, B^{>i}, F, I = i) \\
&\quad (R = (A^{<i}, B^{>i}, F, I))
\end{aligned}$$

$$= \sum_{i=1}^m \frac{1}{m} \cdot \mathbb{I}_{\mathcal{D}_{\text{Disj}}^{\mathbf{N}}}(\Pi_{\text{Disj}} : A_i \mid B_i, A^{<i}, B^{>i}, F)$$

where the last equality is true since conditioned on $(A, B) \sim \mathcal{D}_{\text{Disj}}^{\mathbf{N}}$, all sets A_j, B_j (for $j \in [m]$) are chosen from $\mathcal{D}_{\text{Disj}}^{\mathbf{N}}$ and hence are independent of the “ $I = i$ ” event⁶. Define $\mathbf{A} := (A_1, \dots, A_m)$ and $\mathbf{B} := (B_1, \dots, B_m)$; we can further derive,

$$\begin{aligned}
\mathbb{I}_{\mathcal{D}_{\text{Disj}}^{\mathbf{N}}}(\Pi_{\text{Disj}} : A \mid B, R) &= \sum_{i=1}^m \frac{1}{m} \cdot \mathbb{I}_{\mathcal{D}_{\text{Disj}}^{\mathbf{N}}}(\Pi_{\text{Disj}} : A_i \mid B_i, A^{<i}, B^{>i}, F) \\
&\leq \frac{1}{m} \cdot \sum_{i=1}^m \mathbb{I}_{\mathcal{D}_{\text{Disj}}^{\mathbf{N}}}(\Pi_{\text{Disj}} : A_i \mid A^{<i}, \mathbf{B}, F) \\
&\quad (A_i \perp B^{<i} \mid \mathbf{B}, F \text{ and hence we can apply Fact A.2}) \\
&= \frac{1}{m} \cdot \mathbb{I}_{\mathcal{D}_{\text{Disj}}^{\mathbf{N}}}(\Pi_{\text{Disj}} : \mathbf{A} \mid \mathbf{B}, F) \\
&\quad (\text{chain rule of mutual information, Fact A.1-(4)}) \\
&= \frac{1}{m} \cdot \mathbb{I}_{\mathcal{D}_{\text{Disj}}^{\mathbf{N}}}(\Pi_{\text{Disj}} : \mathcal{S} \mid \mathcal{T}, F) \\
&= \frac{1}{m} \cdot \mathbb{I}_{\mathcal{D}_{\text{SC}}}(\Pi_{\text{SC}} : \mathcal{S} \mid \mathcal{T}, F, \theta = 0)
\end{aligned}$$

where the second last equality is because \mathbf{A} (resp. \mathbf{B}) and \mathcal{S} (resp. \mathcal{T}) determine each other conditioned on F , and last equality is because the distribution of set cover instances and the messages communicated by the players under $\mathcal{D}_{\text{Disj}}^{\mathbf{N}}$ and under $\mathcal{D}_{\text{SC}} \mid \theta = 0$ exactly matches.

Moreover,

$$\begin{aligned}
\mathbb{I}_{\mathcal{D}_{\text{Disj}}^{\mathbf{N}}}(\Pi_{\text{Disj}} : A \mid B, R) &\leq \frac{1}{m} \cdot \mathbb{I}_{\mathcal{D}_{\text{SC}}}(\Pi_{\text{SC}} : \mathcal{S} \mid \mathcal{T}, F, \theta = 0) \\
&\leq \frac{2}{m} \cdot \mathbb{I}_{\mathcal{D}_{\text{SC}}}(\Pi_{\text{SC}} : \mathcal{S} \mid \mathcal{T}, F, \theta) \\
&\quad (\text{by definition of mutual information as } \mathbb{P}(\theta = 0) = 1/2) \\
&\leq \frac{2}{m} \cdot (\mathbb{I}_{\mathcal{D}_{\text{SC}}}(\Pi_{\text{SC}} : \mathcal{S} \mid \mathcal{T}, F) + H(\theta)) \quad (\text{by Fact A.4}) \\
&= \frac{2}{m} \cdot \mathbb{I}_{\mathcal{D}_{\text{SC}}}(\Pi_{\text{SC}} : \mathcal{S} \mid \mathcal{T}, F) + \frac{2}{m} \quad (H(\theta) = 1 \text{ by Fact A.1-(1)}) \\
&\leq \frac{2}{m} \cdot \mathbb{I}_{\mathcal{D}_{\text{SC}}}(\Pi_{\text{SC}} : \mathcal{S} \mid \mathcal{T}) + \frac{2}{m} \\
&\quad (\Pi_{\text{SC}} \perp F \mid \mathcal{S}, \mathcal{T} \text{ and hence we can apply Fact A.3})
\end{aligned}$$

By performing the same exact calculation for $\mathbb{I}_{\mathcal{D}_{\text{Disj}}^{\mathbf{N}}}(\Pi_{\text{Disj}} : B \mid A, R)$, we obtain that,

$$\text{ICost}_{\mathcal{D}_{\text{Disj}}^{\mathbf{N}}}(\pi_{\text{Disj}}) \leq \frac{2}{m} \cdot (\mathbb{I}_{\mathcal{D}_{\text{SC}}}(\Pi_{\text{SC}} : \mathcal{S} \mid \mathcal{T}) + \mathbb{I}_{\mathcal{D}_{\text{SC}}}(\Pi_{\text{SC}} : \mathcal{T} \mid \mathcal{S})) + \frac{4}{m}$$

⁶We point out that this is the exact reason we need to consider information cost of π_{Disj} on $\mathcal{D}_{\text{Disj}}^{\mathbf{N}}$ (instead of $\mathcal{D}_{\text{Disj}}$) as otherwise (A_j, B_j) 's are *not* independent of $I = i$ and hence this equality would not hold.

$$= \frac{2}{m} \cdot \text{ICost}_{\mathcal{D}_{\text{SC}}}(\pi_{\text{SC}}) + \frac{4}{m} = \frac{O(1)}{m} \cdot \text{ICost}_{\mathcal{D}_{\text{SC}}}(\pi_{\text{SC}})$$

where in the last inequality we used the fact that information cost of π_{SC} is at least 1. This finalizes the proof of the lemma. \blacksquare

Recall that in Lemma 3.4, we bound the information cost of π_{Disj} on the distribution $\mathcal{D}_{\text{Disj}}^{\text{N}}$ (as opposed to $\mathcal{D}_{\text{Disj}}$); in the following we prove that this weaker bound is still sufficient for our purpose.

Lemma 3.5. *For any $\delta < 1/2$, any δ -error protocol π_{Disj} for Disj_t on $\mathcal{D}_{\text{Disj}}$ with $\|\pi_{\text{Disj}}\| = 2^{o(t)}$ has*

$$\text{ICost}_{\mathcal{D}_{\text{Disj}}^{\text{N}}}(\pi_{\text{Disj}}) = \Omega(t).$$

By Proposition 2.5, any δ -error protocol for Disj_t (with $\delta < 1/2$) on $\mathcal{D}_{\text{Disj}}$ has $\text{ICost}_{\mathcal{D}_{\text{Disj}}^{\text{Y}}}(\pi_{\text{Disj}}) = \Omega(t)$ (notice again that the information cost is measured on the distribution $\mathcal{D}_{\text{Disj}}^{\text{Y}}$). From this, it is also easy to obtain that $\text{ICost}_{\mathcal{D}_{\text{Disj}}}(\pi_{\text{Disj}}) = \Omega(t)$. However, to prove Lemma 3.5, we need to lower bound the information cost of π_{Disj} under the distribution $\mathcal{D}_{\text{Disj}}^{\text{N}}$.

To achieve this, we can relate the information costs $\text{ICost}_{\mathcal{D}_{\text{Disj}}^{\text{Y}}}(\pi_{\text{Disj}})$ and $\text{ICost}_{\mathcal{D}_{\text{Disj}}^{\text{N}}}(\pi_{\text{Disj}})$ to each other. The goal is to argue that if there is a large discrepancy in the information cost of π_{Disj} on $\mathcal{D}_{\text{Disj}}^{\text{Y}}$ and $\mathcal{D}_{\text{Disj}}^{\text{N}}$, then the information cost of the protocol itself can be used to distinguish between these two cases. We can achieve this goal using an elegant construction of an “information odometer” by [14]; informally speaking, the odometer allows the players to “keep track” of the amount of information revealed in a protocol (i.e., the information cost of the protocol), while incurring a relatively small additional information cost overhead.

Intuitively, we can use the odometer to argue that $\text{ICost}_{\mathcal{D}_{\text{Disj}}^{\text{N}}}(\pi_{\text{Disj}}) = \Theta(\text{ICost}_{\mathcal{D}_{\text{Disj}}^{\text{Y}}}(\pi_{\text{Disj}}))$ as follows: suppose towards a contradiction that $\text{ICost}_{\mathcal{D}_{\text{Disj}}^{\text{N}}}(\pi_{\text{Disj}}) = \tau$ for some $\tau = o(\text{ICost}_{\mathcal{D}_{\text{Disj}}^{\text{Y}}}(\pi_{\text{Disj}}))$ and consider a new protocol π'_{Disj} for Disj on $\mathcal{D}_{\text{Disj}}$ which runs π_{Disj} and the information odometer for π_{Disj} in parallel. Whenever the odometer estimates the information cost of π_{Disj} to be larger than $c \cdot \tau$ (for some sufficiently large constant c), the players terminate the protocol and declare that the answer for Disj is No (as information cost of π_{Disj} on $\mathcal{D}_{\text{Disj}}^{\text{N}}$ is typically not much more than τ , while its information cost on $\mathcal{D}_{\text{Disj}}^{\text{Y}}$ is $\omega(\tau)$). If the cost is not estimated more than $c \cdot \tau$ by the end of the protocol, the players output the same answer as in π_{Disj} . As the information cost of the information odometer itself is bounded by $O(\tau)$, this results in protocol π'_{Disj} to have $\text{ICost}_{\mathcal{D}_{\text{Disj}}}(\pi'_{\text{Disj}}) = o(t)$, a contradiction. This argument was first made explicit in [29].

Lemma 3.6 (Lemma 15 in [29]). *Fix any function F , constants $0 < \varepsilon_1 < \varepsilon_2 < 1/2$, input distribution \mathcal{D} , and define $\mathcal{D}^{\text{N}} := \mathcal{D} \mid F^{-1}(\text{No})$. For every ε_1 -error protocol π for F on \mathcal{D} , there exists an ε_2 -error protocol π' for F on \mathcal{D} such that:*

$$\text{ICost}_{\mathcal{D}}(\pi') = O(\text{ICost}_{\mathcal{D}^{\text{N}}}(\pi) + \log \|\pi\|).$$

We are now ready to prove Lemma 3.5.

Proof of Lemma 3.5. Let π'_{Disj} be any δ' -error protocol for Disj on $\mathcal{D}_{\text{Disj}}$ for $\delta' < 1/2$. We first prove that $\text{ICost}_{\mathcal{D}_{\text{Disj}}}(\pi'_{\text{Disj}}) = \Omega(t)$ using the fact that $\text{ICost}_{\mathcal{D}_{\text{Disj}}^{\text{Y}}}(\pi'_{\text{Disj}}) = \Omega(t)$ as follows:

$$\begin{aligned} \text{ICost}_{\mathcal{D}_{\text{Disj}}}(\pi'_{\text{Disj}}) &= \mathbb{I}_{\mathcal{D}_{\text{Disj}}}(\Pi'_{\text{Disj}} : A \mid B) + \mathbb{I}_{\mathcal{D}_{\text{Disj}}}(\Pi'_{\text{Disj}} : B \mid A) \\ &\geq \mathbb{I}_{\mathcal{D}_{\text{Disj}}}(\Pi'_{\text{Disj}} : A \mid B, \theta) + \mathbb{I}_{\mathcal{D}_{\text{Disj}}}(\Pi'_{\text{Disj}} : B \mid A, \theta) - 2\mathbb{H}(\theta) \end{aligned} \quad (\text{by Fact A.4})$$

$$\begin{aligned}
&\geq \frac{1}{2} \cdot \mathbb{I}_{\mathcal{D}_{\text{Disj}}}(\Pi'_{\text{Disj}} : A \mid B, \theta = 0) + \frac{1}{2} \cdot \mathbb{I}_{\mathcal{D}_{\text{Disj}}}(\Pi'_{\text{Disj}} : B \mid A, \theta = 0) - 2 \\
&\quad \text{(by definition of mutual information and since } \mathbb{H}(\theta) = 1) \\
&= \frac{1}{2} \cdot \left(\mathbb{I}_{\mathcal{D}_{\text{Disj}}^Y}(\Pi'_{\text{Disj}} : A \mid B) + \mathbb{I}_{\mathcal{D}_{\text{Disj}}^Y}(\Pi'_{\text{Disj}} : B \mid A) \right) - 2 \quad (\mathcal{D}_{\text{Disj}}^Y = \mathcal{D}_{\text{Disj}} \mid \theta = 0) \\
&= \frac{1}{2} \cdot \text{ICost}_{\mathcal{D}_{\text{Disj}}^Y}(\pi'_{\text{Disj}}) - 2 = \Omega(t) \quad \text{(by Proposition 2.5)}
\end{aligned}$$

Now suppose towards a contradiction that $\text{ICost}_{\mathcal{D}_{\text{Disj}}^N}(\pi_{\text{Disj}})$ is $o(t)$. We can then apply Lemma 3.6 for the function $F = \text{Disj}$, $\varepsilon_1 = \delta$ and $\varepsilon_2 = \delta' < 1/2$ to obtain a protocol π'_{Disj} with $\text{ICost}_{\mathcal{D}_{\text{Disj}}}(\pi'_{\text{Disj}}) = O(\text{ICost}_{\mathcal{D}^N}(\pi_{\text{Disj}}) + \log \|\pi_{\text{Disj}}\|)$ which is $o(t)$; a contradiction. ■

We now conclude,

Theorem 3. For any constant $\delta < 1/2$, $\alpha = o(\frac{\log n}{\log \log n})$, and $m = \text{poly}(n)$,

$$\text{CC}_{\mathcal{D}_{\text{SC}}}^\delta(\text{SetCover}) = \tilde{\Omega}(mn^{\frac{1}{\alpha}}).$$

Proof. Let $t = \Theta\left(\left(\frac{n}{\log m}\right)^{\frac{1}{\alpha}}\right)$ and suppose towards a contradiction that there exists a δ -error protocol π_{SC} for **SetCover** on the distribution \mathcal{D}_{SC} with $\|\pi_{\text{SC}}\| = o(mt)$; by Proposition 2.4, $\text{ICost}_{\mathcal{D}_{\text{SC}}}(\pi_{\text{SC}}) = o(mt)$ also. By Lemma 3.4, this implies that there exists a $(\delta + o(1))$ -error protocol π_{Disj} for **Disj** on the distribution $\mathcal{D}_{\text{Disj}}$ such that $\text{ICost}_{\mathcal{D}_{\text{Disj}}^N}(\pi_{\text{Disj}}) = o(t)$, and $\|\pi_{\text{Disj}}\| = o(mt) \leq 2^{o(t)}$ (since $m = \text{poly}(n)$ and $\alpha = o(\frac{\log n}{\log \log n})$). However, this is in contradiction with Lemma 3.5, implying that $\|\pi_{\text{SC}}\| = \Omega(mt)$, hence proving the theorem. ■

As a corollary of Theorem 3, we have that the space complexity of any α -approximation streaming algorithm for set cover that uses $\text{polylog}(n)$ passes on *adversarial streams* is $\tilde{\Omega}(mn^{\frac{1}{\alpha}})$. In the next section, we extend this result to random arrival streams and complete the proof of Theorem 1.

3.3 Proof of Theorem 1

The distribution \mathcal{D}_{SC} used in the previous section is quite “adversarial” and as such is not suitable for proving the lower bound for random arrival streams. In order to prove the lower bound in Theorem 1 for random arrival streams, we need to relax the adversarial partitioning of the sets in the distribution \mathcal{D}_{SC} to a randomized partition.

Distribution $\mathcal{D}_{\text{SC}}^{\text{rnd}}$. A random partitioning of the distribution \mathcal{D}_{SC}

- Sample the collections $(\mathcal{S}, \mathcal{T}) \sim \mathcal{D}_{\text{SC}}$.
- Assign each set in $\mathcal{S} \cup \mathcal{T}$ to Alice w.p. $1/2$ and the remainings to Bob.

We show that even this seemingly easier distribution still captures all the “hardness” of distribution \mathcal{D}_{SC} . Formally,

Lemma 3.7. For any constant $\delta < 1/4$, $\alpha = o(\frac{\log n}{\log \log n})$, and $m = \text{poly}(n)$,

$$\text{CC}_{\mathcal{D}_{\text{SC}}^{\text{rnd}}}^\delta(\text{SetCover}) = \tilde{\Omega}(mn^{\frac{1}{\alpha}})$$

Proof. Let $\mathcal{S} = \{S_1, \dots, S_m\}$ and $\mathcal{T} = \{T_1, \dots, T_m\}$ be the collections of sets sampled from \mathcal{D}_{SC} in the distribution $\mathcal{D}_{\text{SC}}^{\text{rnd}}$. For a sampled instance in $\mathcal{D}_{\text{SC}}^{\text{rnd}}$, we say that the index $i \in [m]$ is *good* iff S_i is given to one player and T_i to another. Let $G \subseteq [m]$ be the collection of all good indices. The index i^* is chosen independent of the random partitioning in $\mathcal{D}_{\text{SC}}^{\text{rnd}}$, and hence the probability that $i^* \in G$ is exactly $|G|/m$. Let \mathcal{E} denote the event that $|G| \geq m/2 - o(m)$ and $i^* \in G$. We have,

$$\begin{aligned} \mathbb{P}(\mathcal{E}) &= \mathbb{P}(|G| \geq m/2 - o(m)) \cdot \mathbb{P}(i^* \in G \mid |G| \geq m/2 - o(m)) \\ &\geq \mathbb{P}(|G| \geq (1 - o(1)) \cdot \mathbb{E}[G]) \cdot \frac{1 - o(1)}{2} \geq (1 - o(1)) \cdot \frac{1}{2} \end{aligned}$$

where the last inequality is by Chernoff bound. Now fix a δ -error protocol π_{SC} for **SetCover** on the distribution $\mathcal{D}_{\text{SC}}^{\text{rnd}}$. Then,

$$\mathbb{P}(\pi_{\text{SC}} \text{ errs} \mid \mathcal{E}) \leq \frac{\mathbb{P}(\pi_{\text{SC}} \text{ errs})}{\mathbb{P}(\mathcal{E})} \leq 2\delta + o(1) \quad (3)$$

This in particular implies that there exists a set $G^* \subseteq [n]$ with $|G^*| \geq m/2 - o(m)$, such that conditioned on the set of good indices being G^* and conditioned on $i^* \in G^*$, the probability that π_{SC} errs is at most $2\delta + o(1)$. Note that conditioned on the aforementioned events, the index i^* is chosen from G^* uniformly at random. This implies that the distribution of the input given to Alice and Bob limited to the sets in G^* matches the distribution \mathcal{D}_{SC} (with the number of the sets being $2 \cdot |G^*|$ instead of $2m$). We can then use this to embed an instance of **SetCover** over the distribution \mathcal{D}_{SC} into the sets G^* and obtain a protocol π'_{SC} for \mathcal{D}_{SC} .

More formally, the protocol π'_{SC} works as follows: Given an instance $(\mathcal{S}', \mathcal{T}')$ sampled from \mathcal{D}_{SC} (with $|\mathcal{S}'| = |\mathcal{T}'| = |G^*|$), Alice and Bob use public coins to complete their input (i.e., increase the number of the sets to $2m$) by sampling from the distribution $\mathcal{D}_{\text{SC}}^{\text{rnd}}$ conditioned on G^* (this is possible without any communication as the sets outside G^* are sampled independent of the sets in G^*). The players then run the protocol π_{SC} on this new instance and return the same answer as this protocol. As the distribution of the **SetCover** instances sampled in the protocol π'_{SC} matches the distribution $\mathcal{D}_{\text{SC}}^{\text{rnd}}$ conditioned on G^* and $i^* \in G^*$, by Eq (3), the probability that π'_{SC} errs is at most $2\delta + o(1)$. Since $\delta < 1/4$, we obtain a δ' -error protocol for **SetCover** on the distribution \mathcal{D}_{SC} with $2|G^*| = \Theta(m)$ sets and universe of size n , for a constant $\delta' < 1/2$. Consequently, by Theorem 3, $\|\pi_{\text{SC}}\| = \|\pi'_{\text{SC}}\| = \tilde{\Omega}(|G^*| \cdot n^{\frac{1}{\alpha}}) = \tilde{\Omega}(mn^{\frac{1}{\alpha}})$, proving the lemma. ■

We are now ready to prove Theorem 1.

Proof of Theorem 1. Fix a p -pass s -space streaming algorithm \mathcal{A} for the set cover problem over random arrival streams that outputs an α -approximation w.p. at least $1 - \delta$ for $\delta < 1/4$. One can easily turn \mathcal{A} into a δ -error protocol for **SetCover** on the distribution $\mathcal{D}_{\text{SC}}^{\text{rnd}}$: Alice and Bob take a random permutation of their inputs and then treat their combined input as a set stream and run \mathcal{A} on that. The random partitioning of the input plus the random permutation taken by the players ensure that the constructed stream is a random permutation of the input sets. Consequently, this protocol is a δ -error protocol for **SetCover** on $\mathcal{D}_{\text{SC}}^{\text{rnd}}$ that uses $O(p \cdot s)$ bits of communication. Since $\delta < 1/4$, by Lemma 3.7, $p \cdot s = \tilde{\Omega}(mn^{\frac{1}{\alpha}})$, proving the theorem. ■

3.4 An α -Approximation Algorithm for the Streaming Set Cover Problem

In this section, we prove the optimality of the lower bound in Theorem 1 by establishing a matching upper bound (i.e. Theorem 2). As stated earlier, our algorithm is a simple modification of the algorithm of [32]. In particular, we obtain our improved algorithm by using a one-shot pruning

step as opposed to the iterative pruning of [32], and employing a more careful element sampling (compare the bounds in Lemma 3.12 in this paper with Lemma 2.5 in [32]).

In the following, we assume that we are given a value $\widetilde{\text{opt}}$ which is a $(1 + \varepsilon)$ -approximation of opt , i.e., the optimal solution size of the given instance. This is without loss of generality as we can run the algorithm in parallel for $O(\log n/\varepsilon)$ guesses for $\widetilde{\text{opt}} \in [1, n]$ and return the smallest computed set cover among all parallel runs.

The general idea behind the algorithm is as follows: we know that $\widetilde{\text{opt}}$ sets are enough to cover the whole universe $[n]$; hence, if we find a $(1 - \rho)$ -approximate k -cover of the input sets for the parameter $k = \widetilde{\text{opt}}$ and $\rho = 1/n^{1/\alpha}$, we can reduce the number of uncovered elements by a factor of $n^{1/\alpha}$. Repeating this process α times then results in a collection of at most $\alpha \cdot \widetilde{\text{opt}}$ sets that covers the whole universe, i.e., an α -approximate set cover. It is worth mentioning that this is the general principle behind most (but not all) streaming algorithms for set cover, see, e.g. [9, 23, 32, 47].

Notice that we can readily use the maximum coverage streaming algorithms of [9, 42] as a sub-routine to find the approximate k -cover above; however, doing so would result in a sub-optimal algorithm for set cover as these algorithms have space dependence of (at least) $\Omega(m/\rho^2) = \Omega(mn^{2/\alpha})$ (even ignoring the dependence on k , i.e., $\widetilde{\text{opt}}$). In fact, as we prove in the next section (see Result 2), any $(1 - \rho)$ -approximate k -cover algorithm needs $\Omega(m/\rho^2)$ space in general. To bypass this, we crucially use the fact that the aforementioned maximum coverage instances have the additional property that the optimal answer is the whole universe and hence the element sampling technique of [32] (and similar ones in [9, 42]) can be improved for this special case. We now provide the formal description of the algorithm.

Algorithm 1. An α -approximation algorithm for the streaming set cover problem.

Input. A stream $\mathcal{S} = (S_1, \dots, S_m)$ of subsets of $[n]$, and a $(1 + \varepsilon)$ -approximation $\widetilde{\text{opt}}$ of $\text{opt}(\mathcal{S})$.

Output. A collection of $(1 + \varepsilon) \cdot \alpha \cdot \widetilde{\text{opt}}$ sets that cover the universe.

-
1. Let $U \leftarrow [n]$ and $\text{SOL} \leftarrow \emptyset$.
 2. Make a single pass over the stream and if $|S_i \cap U| \geq n/(\varepsilon \cdot \widetilde{\text{opt}})$, then:
 - (a) $\text{SOL} \leftarrow \text{SOL} \cup \{i\}$ and $U \leftarrow U \setminus S_i$.
 3. For $j = 1$ to α iterations:
 - (a) Let U_{smp} be a subset of U chosen by picking each element independently and w.p. $p = 16 \cdot \widetilde{\text{opt}} \cdot \log m / n^{1-1/\alpha}$.
 - (b) Make a single pass over the stream and for all $i \in [m]$, store $S'_i = S_i \cap U_{\text{smp}}$ in the memory.
 - (c) Find an optimal set cover OPT' of the instance (S'_1, \dots, S'_m) and let $\text{SOL} \leftarrow \text{SOL} \cup \text{OPT}'$.
 - (d) Make another pass over the stream and let $U_{\text{smp}} \leftarrow U_{\text{smp}} \setminus \bigcup_{i \in \text{OPT}'} S_i$.
 4. Return SOL as a set cover of the input instance.

We start by bounding the space requirement of Algorithm 1.

Lemma 3.8. *Algorithm 1 requires $\widetilde{O}(mn^{1/\alpha}/\varepsilon + n)$ space w.p. at least $1 - 1/m^2$.*

Proof. It is easy to see that maintaining SOL and U requires, respectively, $O(m)$ and $O(n)$ space.

In the following, we analyze the space required for storing the sets (S'_1, \dots, S'_m) . After the first pass of the algorithm, no set contains more than $n/(\varepsilon \cdot \widetilde{\text{opt}})$ elements in U . Fix a set $S_i \in \mathcal{S}$; we have,

$$\begin{aligned} \mathbb{E} |S_i \cap U_{\text{smp}}| &= |S_i| \cdot p \leq n/(\varepsilon \cdot \widetilde{\text{opt}}) \cdot \left(16 \cdot \widetilde{\text{opt}} \cdot \log m / n^{1-1/\alpha}\right) \\ &= 16 \cdot n^{1/\alpha} \cdot \log m / \varepsilon \end{aligned}$$

Hence, by Chernoff bound, w.p. $1 - 1/m^3$, $|S_i \cap U_{\text{smp}}| = \widetilde{O}(n^{1/\alpha}/\varepsilon)$. The final bound now follows from this and a union bound on all m sets in \mathcal{S} . \blacksquare

Remark 3.9. *One can make the space requirement of Algorithm 1 deterministic by terminating the algorithm whenever it attempts to use a memory more than the bounds in Lemma 3.8. As this event happens with negligible probability, the correctness of the algorithm can be argued exactly the same.*

The following two lemmas establish the correctness of the algorithm.

Lemma 3.10. *Algorithm 1 picks at most $(\alpha + \varepsilon) \cdot \widetilde{\text{opt}}$ sets in SOL.*

Proof. It is immediate to see that in the first pass, the algorithm picks at most $\varepsilon \cdot \widetilde{\text{opt}}$ sets as otherwise U would be empty. Moreover, in each subsequent α iterations, the algorithm picks at most $\widetilde{\text{opt}}$ sets since (S'_1, \dots, S'_m) has a set cover of size at most $\widetilde{\text{opt}}$ (as the original instance had a set cover of size $\leq \text{opt}$). \blacksquare

Lemma 3.11. *The set SOL computed by Algorithm 1 is a feasible set cover of $[n]$ w.p. $1 - 1/m$.*

To prove Lemma 3.11, we use the following property of element sampling that first appeared in [23] (similar ideas also appear in [32, 42]); for completeness we provide a self-contained proof of this lemma here.

Lemma 3.12. *Let $0 < \rho < 1$ be a parameter and $\mathcal{S} = (S_1, \dots, S_m)$ be a collection of m subsets of $[n]$ with $\text{opt}(\mathcal{S}) \leq k$. Suppose U_{smp} is a subset of $[n]$ obtained by picking each element independently and w.p. $p \geq 16 \cdot k \cdot \log m / (\rho \cdot n)$; then, w.p. $1 - 1/m^2$, any collection of k sets in \mathcal{S} that covers U_{smp} entirely also covers at least $(1 - \rho) \cdot n$ elements in $[n]$.*

Proof. Fix a collection \mathcal{C} of k subsets in \mathcal{S} that covers less than $(1 - \rho) \cdot n$ elements in $[n]$. The probability that this collection covers U_{smp} entirely is equal to the probability that none of the $\rho \cdot n$ elements in $[n]$ that are not appearing in \mathcal{C} are sampled in U_{smp} . Hence,

$$\mathbb{P}(\mathcal{C} \text{ covers } U_{\text{smp}}) \leq (1 - p)^{\rho \cdot n} \leq \exp(- (16 \cdot k \cdot \log m / (\rho \cdot n)) \cdot (\rho \cdot n)) \leq 1/m^{8k}$$

Taking a union bound over all $\binom{m}{k} \leq m^k$ possible choices for \mathcal{C} finalizes the result. \blacksquare

Proof of Lemma 3.11. In each of the α iterations, Algorithm 1 implements the sampling in Lemma 3.12 with the parameters $k = \widetilde{\text{opt}}$, and $\rho = n^{-1/\alpha}$. Hence, after each iteration, the number of uncovered elements in U reduces to $|U|/n^{1/\alpha}$ w.p. $1 - 1/m^2$. Consequently, by taking a union bound over the $\alpha \leq m$ iterations, after the α iterations, number of uncovered elements reduces to less than 1, hence proving the lemma. \blacksquare

We now conclude the proof of Theorem 2.

Proof of Theorem 2. We can run Algorithm 1 in parallel for $O(\log n/\varepsilon)$ possible guesses for $\widetilde{\text{opt}}$. By Lemma 3.8, the space requirement of this algorithm is $\widetilde{O}(1/\varepsilon) \cdot \widetilde{O}(mn^{1/\alpha}/\varepsilon + n)$ as desired. Moreover, consider the guess: $\text{opt} \leq \widetilde{\text{opt}} \leq (1 + \varepsilon) \cdot \widetilde{\text{opt}}$. For this choice, we can apply Lemma 3.10 and Lemma 3.11 and obtain that the returned solution is an $(\alpha + O(\varepsilon))$ -approximation of the optimal set cover. Since the algorithm can make sure that the returned solution is always feasible, returning the smallest set cover among all guesses for $\widetilde{\text{opt}}$ then ensures that the returned answer is an $(\alpha + O(\varepsilon))$ approximation. Re-parameterizing ε by a constant factor, finalizes the proof. \blacksquare

4 The Space-Approximation Tradeoff for Maximum Coverage

In this section, we prove a space-approximation tradeoff for the maximum coverage problem.

Theorem 4. *For any $\varepsilon = \omega(1/\sqrt{n})$, $m = \text{poly}(n)$, and $p \geq 1$, any randomized algorithm that can make p passes over any collection of m subsets of $[n]$ presented in a random order stream and outputs a $(1 - \varepsilon)$ -approximation to the optimal value of the maximum coverage problem for $k = 2$ with a sufficiently large constant probability (over the randomness of both the stream order and the algorithm) must use $\Omega(m/(\varepsilon^2 \cdot p))$ space.*

Similar to previous section, we prove Theorem 4 by considering the communication complexity of the maximum coverage problem: Fix a (sufficiently large) n , $\varepsilon = \omega(1/\sqrt{n})$ and $m = \text{poly}(n)$; **MaxCover** refers to the communication problem of $(1 - \varepsilon)$ -approximating the optimal value of the maximum coverage problem with $2m$ sets defined over the universe $[n]$ and parameter $k = 2$, in the two-player communication model.

Our lower bound for **MaxCover** is obtained by reducing this problem to multiple instances of the *gap-hamming-distance* problem via a similar distribution as \mathcal{D}_{SC} (using an additional simple gadget). In the following, we first introduce the gap-hamming-distance problem and prove a useful lemma on its information complexity on particular distributions required for our reduction, and then describe a hard distribution for **MaxCover** based on this and finalize the proof of Theorem 4.

4.1 The Gap-Hamming-Distance Problem

The *gap-hamming-distance* (GHD) problem is defined as follows. Fix an integer $t \geq 1$; in GHD_t , Alice is given a set $A \subseteq [t]$, Bob is given a set $B \subseteq [t]$ and the goal is to output:

$$\text{GHD}(A, B) = \begin{cases} \text{Yes} & \Delta(A, B) \geq t/2 + \sqrt{t} \\ \text{No} & \Delta(A, B) \leq t/2 - \sqrt{t} \\ \star & \text{otherwise} \end{cases}$$

where \star means that the answer can be arbitrary; here $\Delta(A, B)$ denotes the hamming distance between A and B , i.e., the size of the symmetric difference of A and B .

This problem was originally introduced by [35] and has been studied extensively in the literature (see [16] and references therein). We use the following result on the information complexity of this problem proven in [12]⁷.

Lemma 4.1 ([12]). *Let \mathcal{U} be the uniform distribution on pairs of subsets of $[t]$ (chosen independently); there exists an absolute constant $\delta > 0$ such that*

$$\text{IC}_{\mathcal{U}}^{\delta}(\text{GHD}) = \Omega(t).$$

⁷Technically speaking, [12] bounds the *external* information complexity of GHD as opposed to its *internal* information complexity used in our paper. However, since the distribution in Lemma 4.1 is a product distribution, these two quantities are equal and hence we simply state the bound for the internal information complexity.

For our purpose, we need to consider the following distribution \mathcal{D}_{GHD} for GHD instead of the uniform distribution. Let $a, b \in [t]$ be two parameters to be determined later⁸. Define:

- $\mathcal{D}_{\text{GHD}}^Y$ as the distribution of instances $(A, B) \sim \mathcal{U} \mid \Delta(A, B) \geq t/2 + \sqrt{t}, |A| = a, |B| = b$.
- $\mathcal{D}_{\text{GHD}}^N$ as the distribution of instances $(A, B) \sim \mathcal{U} \mid \Delta(A, B) \leq t/2 - \sqrt{t}, |A| = a, |B| = b$.
- $\mathcal{D}_{\text{GHD}} := \frac{1}{2} \cdot \mathcal{D}_{\text{GHD}}^Y + \frac{1}{2} \cdot \mathcal{D}_{\text{GHD}}^N$.

We use Lemma 4.1 to prove the following result on the information cost of δ -error protocols on the distribution \mathcal{D}_{GHD} , which could be independently useful also. The proof is deferred to Appendix B.

Lemma 4.2. *Let $\delta > 0$ be a sufficiently small constant and π_{GHD} be a δ -error protocol for GHD_t on \mathcal{D}_{GHD} with $\|\pi_{\text{GHD}}\| = 2^{o(t)}$; then, $\text{ICost}_{\mathcal{D}_{\text{GHD}}}(\pi_{\text{GHD}}) = \Omega(t)$.*

4.2 Communication Complexity of MaxCover

We are now ready to prove a lower bound on the communication complexity of the MaxCover problem. To do so, we propose the following distribution.

Distribution \mathcal{D}_{MC} . A hard input distribution for MaxCover.

Notation. Let $t_1 := 1/\varepsilon^2$, $t_2 := 10 \cdot t_1$, $U_1 := [t_1]$ and $U_2 := [t_1 + 1, t_1 + t_2]$.

- For each $i \in [m]$:
 - Let $(A_i, B_i) \sim \mathcal{D}_{\text{GHD}}^N$ for GHD_{t_1} on the universe U_1 .
 - Create $C_i, D_i \subseteq U_2$, by assigning each element in U_2 w.p. $1/2$ to C_i and o.w. to D_i .
 - Let $S_i := A_i \cup C_i$ and $T_i := B_i \cup D_i$.
- Pick $\theta \in_R \{0, 1\}$ uniformly at random. If $\theta = 0$, do nothing, otherwise:
 - Sample $i^* \in_R [m]$ uniformly at random.
 - Resample $(A_{i^*}, B_{i^*}) \sim \mathcal{D}_{\text{GHD}}^Y$ for GHD_{t_1} and redefine S_{i^*} and T_{i^*} as before using the new pair (A_{i^*}, B_{i^*}) (do not change C_i and D_i).
- Let the input to Alice and Bob be $\mathcal{S} := \{S_i\}_{i \in [m]}$ and $\mathcal{T} := \{T_i\}_{i \in [m]}$, respectively.

Define $\text{opt}(\mathcal{S}, \mathcal{T})$ as the value of the optimal solution of the maximum coverage problem (for the parameter $k = 2$) for the instance $(\mathcal{S}, \mathcal{T})$. We wish to argue that $\text{opt}(\mathcal{S}, \mathcal{T})$ differs by a $(1 \pm \varepsilon)$ factor depending on the choice of θ in the distribution and hence any $(1 - \varepsilon)$ approximation algorithm for maximum coverage on this distribution needs to determine the value of θ .

Lemma 4.3. *Assuming $\varepsilon = o(1/\log n)$, there exists a fixed $\tau \in [n]$ such that for any instance $(\mathcal{S}, \mathcal{T}) \sim \mathcal{D}_{\text{MC}}$:*

$$\begin{aligned} \mathbb{P}(\text{opt}(\mathcal{S}, \mathcal{T}) \geq (1 + \Theta(\varepsilon)) \cdot \tau \mid \theta = 1) &= 1 - o(1) \\ \mathbb{P}(\text{opt}(\mathcal{S}, \mathcal{T}) \leq (1 - \Theta(\varepsilon)) \cdot \tau \mid \theta = 0) &= 1 - o(1) \end{aligned}$$

⁸The values of a and b are not important for our purpose and are hence only determined in the proof of Lemma 4.2.

Proof. We first prove that, any $(1 - \varepsilon)$ -approximate 2-cover in this distribution always has to pick a pair of (S_i, T_i) sets (for some $i \in [m]$). This is achieved by considering the projection of the sets on the universe U_2 .

Claim 4.4. *W.p. $1 - o(1)$:*

(a) *For any $i \in [m]$, $|S_i \cup T_i| \geq t_2$.*

(b) *For any $i \neq j \in [m]$, for any $Z_i \in \{S_i, T_i\}$, and $Z_j \in \{S_j, T_j\}$, $|Z_i \cup Z_j| \leq (3/4 + 0.2) \cdot t_2$.*

Proof. Part (a) follows immediately from the fact that U_2 is partitioned between S_i and T_i , and that $|U_2| = t_2$. We now prove Part (b). To do so, we prove that $Z_i \cup Z_j$ can only cover (essentially) $3/4$ fraction of U_2 w.h.p and since the rest of $Z_i \cup Z_j$ is a subset of U_1 with $|U_1| \leq 0.1 \cdot t_2$, we get the final result.

For any element $e \in U_2$, define an indicator random variable $X_e \in \{0, 1\}$ whereby $X_e = 1$ iff $e \in Z_i \cup Z_j$. Since $i \neq j$, the elements in Z_i and Z_j that are in U_2 are chosen independent of each other, and hence $\mathbb{P}(X_e = 1) = 1 - (1/2)^2 = 3/4$. Define $X := \sum_{e \in U_2} X_e$; we have $\mathbb{E}[X] = 3/4 \cdot t_2$ and since X_e variables are independent, by Chernoff bound, $\mathbb{P}(X \geq \mathbb{E}[X] + 0.1 \cdot t_2) \leq \exp(-c \cdot t_2) = o(1/m^2)$ (as $t_2 = \omega(\log n)$ and $m = \text{poly}(n)$). The final result now follows from a union bound on all possible $(\leq (2m)^2)$ pairs. ■

Now consider a pair (S_i, T_i) for some $i \in [m]$ and note that $|S_i \cup T_i| = |U_2| + |A_i \cup B_i| = t_2 + |A_i \cup B_i|$; hence we can simply focus on $A_i \cup B_i \subseteq U_1$ part of $S_i \cup T_i$. Moreover, we have that,

$$|A_i \cup B_i| = \frac{1}{2} \cdot (|A_i| + |B_i| + \Delta(A_i, B_i)) = \frac{1}{2} \cdot (a + b + \Delta(A_i, B_i))$$

where we used the fact that in the distribution \mathcal{D}_{GHD} , $|A_i| = a$ and $|B_i| = b$ always.

Consequently, whenever $(A_i, B_i) \sim \mathcal{D}_{\text{GHD}}^{\text{N}}$, we have,

$$|S_i \cup T_i| = t_2 + |A_i \cup B_i| \leq t_2 + (a + b)/2 + t_1/4 - \sqrt{t_1}/2 = (1 - \Theta(\varepsilon)) \cdot \tau$$

for $\tau := t_2 + (a + b)/2 + t_1/4$. Similarly, whenever $(A_i, B_i) \sim \mathcal{D}_{\text{GHD}}^{\text{Y}}$,

$$|S_i \cup T_i| \geq t_2 + |A_i \cup B_i| \geq t_2 + (a + b)/2 + t_1/4 + \sqrt{t_1}/2 = (1 + \Theta(\varepsilon)) \cdot \tau$$

Combining these bounds with Claim 4.4 finalizes the proof. ■

Having proved Lemma 4.3, we can use any $(1 - \varepsilon)$ -approximation protocol for MaxCover to determine the parameter θ in the distribution \mathcal{D}_{MC} (by a simple re-parametrizing of the ε by a constant factor). This allows us to prove the following lemma. The proof is essentially identical to that of Lemma 3.4 in Section 3.2 and is provided only for the sake of completeness.

Lemma 4.5. *Let π_{MC} be a δ -error protocol for MaxCover on \mathcal{D}_{MC} . There exists a $(\delta + o(1))$ -protocol π_{GHD} for GHD_{t_1} on the distribution \mathcal{D}_{GHD} such that:*

1. $\text{ICost}_{\mathcal{D}_{\text{GHD}}^{\text{N}}}(\pi_{\text{GHD}}) = \frac{O(1)}{m} \cdot \text{ICost}_{\mathcal{D}_{\text{MC}}}(\pi_{\text{MC}})$.
2. $\|\pi_{\text{GHD}}\| = \|\pi_{\text{MC}}\|$.

Proof. We design the protocol π_{GHD} as follows:

Protocol π_{GHD} . The protocol for solving GHD_{t_1} using a protocol π_{MC} for **MaxCover**.

Input: An instance $(A, B) \sim \mathcal{D}_{\text{GHD}}$. **Output:** $\text{GHD}(A, B)$.

1. Using public randomness, the players sample an index $i^* \in_R [m]$.
2. Using public randomness, the players sample the sets $A^{<i^*}$ and $B^{>i^*}$ each from $\mathcal{D}_{\text{GHD}}^{\text{N}}$ independently.
3. Using private randomness, Alice samples the sets $A^{>i^*}$ such that $(A_j, B_j) \sim \mathcal{D}_{\text{GHD}}^{\text{N}}$ (for all $j > i^*$); similarly Bob samples the sets $B^{<i^*}$.
4. Using public randomness the players sample the sets (C_i, D_i) for all $i \in [m]$ from a (distinct) universe U_2 the same as distribution \mathcal{D}_{MC} .
5. The players construct the collections $\mathcal{S} := \{S_1, \dots, S_m\}$ and $\mathcal{T} := \{T_1, \dots, T_m\}$ by setting $S_i := A_i \cup C_i$ and $T_i := B_i \cup D_i$ (exactly as in distribution \mathcal{D}_{MC}).
6. The players solve the **MaxCover** instance using π_{MC} and output **No** iff π_{MC} estimates $\text{opt}(\mathcal{S}, \mathcal{T}) \leq \tau$ (for the parameter τ in Lemma 4.3) and **Yes** otherwise.

It is easy to see that the distribution of instances $(\mathcal{S}, \mathcal{T})$ created in the protocol π_{MC} matches the distribution \mathcal{D}_{MC} for **MaxCover** exactly, and hence by Lemma 4.3, π_{GHD} is a $(\delta + o(1))$ -error protocol for **GHD** in the distribution \mathcal{D}_{MC} . The bound on the communication cost of π_{MC} is also immediate; in the following we bound the information cost of π_{MC} for (A, B) sampled from $\mathcal{D}_{\text{GHD}}^{\text{N}}$.

Let I be a random variable for i^* and R be the set of public randomness used by the players. By Claim 2.3,

$$\text{ICost}_{\mathcal{D}_{\text{GHD}}^{\text{N}}}(\pi_{\text{GHD}}) = \mathbb{I}_{\mathcal{D}_{\text{GHD}}^{\text{N}}}(\Pi_{\text{GHD}} : A \mid B, R) + \mathbb{I}_{\mathcal{D}_{\text{GHD}}^{\text{N}}}(\Pi_{\text{GHD}} : B \mid A, R)$$

We now bound the first term in the RHS above (the second term can be bounded exactly the same). In the following, let \mathbf{C} and \mathbf{D} denote the vector of random variables for C_i 's and D_i 's, respectively.

$$\begin{aligned} \mathbb{I}_{\mathcal{D}_{\text{GHD}}^{\text{N}}}(\Pi_{\text{GHD}} : A \mid B, R) &= \mathbb{I}_{\mathcal{D}_{\text{GHD}}^{\text{N}}}(\Pi_{\text{GHD}} : A \mid B, R, I) \quad (I \text{ is chosen using public randomness}) \\ &= \mathbb{E}_i \left[\mathbb{I}_{\mathcal{D}_{\text{GHD}}^{\text{N}}}(\Pi_{\text{GHD}} : A_i \mid B_i, A^{<i}, B^{>i}, \mathbf{C}, \mathbf{D}, I = i) \right] \\ &\quad (R = (A^{<i}, B^{>i}, \mathbf{C}, \mathbf{D}, I)) \\ &= \sum_{i=1}^m \frac{1}{m} \cdot \mathbb{I}_{\mathcal{D}_{\text{GHD}}^{\text{N}}}(\Pi_{\text{GHD}} : A_i \mid B_i, A^{<i}, B^{>i}, \mathbf{C}, \mathbf{D}) \end{aligned}$$

where the last equality is because conditioned on $(A, B) \sim \mathcal{D}_{\text{GHD}}^{\text{N}}$, all sets A_j, B_j (for $j \in [m]$) are chosen from $\mathcal{D}_{\text{GHD}}^{\text{N}}$ and hence are independent of the “ $I = i$ ” event. We can further derive,

$$\begin{aligned} \mathbb{I}_{\mathcal{D}_{\text{GHD}}^{\text{N}}}(\Pi_{\text{GHD}} : A \mid B, R) &= \sum_{i=1}^m \frac{1}{m} \cdot \mathbb{I}_{\mathcal{D}_{\text{GHD}}^{\text{N}}}(\Pi_{\text{GHD}} : A_i \mid B_i, A^{<i}, B^{>i}, \mathbf{C}, \mathbf{D}) \\ &\leq \frac{1}{m} \cdot \sum_{i=1}^m \mathbb{I}_{\mathcal{D}_{\text{GHD}}^{\text{N}}}(\Pi_{\text{GHD}} : A_i \mid A^{<i}, \mathbf{B}, \mathbf{C}, \mathbf{D}) \\ &\quad (A_i \perp B^{<i} \mid \mathbf{B}, \mathbf{C}, \mathbf{D} \text{ and hence we can apply Fact A.2}) \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{m} \cdot \mathbb{I}_{\mathcal{D}_{\text{GHD}}^{\text{N}}} (\Pi_{\text{GHD}} : \mathbf{A} \mid \mathbf{B}, \mathbf{C}, \mathbf{D}) \\
&\quad (\text{chain rule of mutual information, Fact A.1-(4)}) \\
&= \frac{1}{m} \cdot \mathbb{I}_{\mathcal{D}_{\text{GHD}}^{\text{N}}} (\Pi_{\text{GHD}} : \mathcal{S} \mid \mathcal{T}, \mathbf{C}, \mathbf{D}) \\
&= \frac{1}{m} \cdot \mathbb{I}_{\mathcal{D}_{\text{MC}}} (\Pi_{\text{MC}} : \mathcal{S} \mid \mathcal{T}, \mathbf{C}, \mathbf{D}, \theta = 0)
\end{aligned}$$

where the second last equality is because \mathbf{A} (resp. \mathbf{B}) and \mathcal{S} (resp. \mathcal{T}) determine each other conditioned on \mathbf{C} and \mathbf{D} , and last equality is because the distribution of maximum coverage instances and the messages communicated by the players under $\mathcal{D}_{\text{GHD}}^{\text{N}}$ and under $\mathcal{D}_{\text{MC}} \mid \theta = 0$ exactly matches.

Moreover,

$$\begin{aligned}
\mathbb{I}_{\mathcal{D}_{\text{GHD}}^{\text{N}}} (\Pi_{\text{GHD}} : \mathbf{A} \mid \mathbf{B}, \mathbf{R}) &\leq \frac{1}{m} \cdot \mathbb{I}_{\mathcal{D}_{\text{MC}}} (\Pi_{\text{MC}} : \mathcal{S} \mid \mathcal{T}, \mathbf{C}, \mathbf{D}, \theta = 0) \\
&\leq \frac{2}{m} \cdot \mathbb{I}_{\mathcal{D}_{\text{MC}}} (\Pi_{\text{MC}} : \mathcal{S} \mid \mathcal{T}, \mathbf{C}, \mathbf{D}, \theta) \\
&\quad (\text{by definition of mutual information as } \mathbb{P}(\theta = 0) = 1/2) \\
&\leq \frac{2}{m} \cdot (\mathbb{I}_{\mathcal{D}_{\text{MC}}} (\Pi_{\text{MC}} : \mathcal{S} \mid \mathcal{T}, \mathbf{C}, \mathbf{D}) + H(\theta)) \quad (\text{by Fact A.4}) \\
&= \frac{2}{m} \cdot \mathbb{I}_{\mathcal{D}_{\text{MC}}} (\Pi_{\text{MC}} : \mathcal{S} \mid \mathcal{T}, \mathbf{C}, \mathbf{D}) + \frac{2}{m} \quad (H(\theta) = 1 \text{ by Fact A.1-(1)}) \\
&\leq \frac{2}{m} \cdot \mathbb{I}_{\mathcal{D}_{\text{MC}}} (\Pi_{\text{MC}} : \mathcal{S} \mid \mathcal{T}) + \frac{2}{m} \\
&\quad (\Pi_{\text{MC}} \perp \mathbf{C}, \mathbf{D} \mid \mathcal{S}, \mathcal{T} \text{ and hence we can apply Fact A.3})
\end{aligned}$$

By performing the same exact calculation for $\mathbb{I}_{\mathcal{D}_{\text{GHD}}^{\text{N}}} (\Pi_{\text{GHD}} : \mathbf{B} \mid \mathbf{A}, \mathbf{R})$, we obtain that,

$$\begin{aligned}
\text{ICost}_{\mathcal{D}_{\text{GHD}}^{\text{N}}} (\pi_{\text{GHD}}) &\leq \frac{2}{m} \cdot (\mathbb{I}_{\mathcal{D}_{\text{MC}}} (\Pi_{\text{MC}} : \mathcal{S} \mid \mathcal{T}) + \mathbb{I}_{\mathcal{D}_{\text{MC}}} (\Pi_{\text{SC}} : \mathcal{T} \mid \mathcal{S})) + \frac{4}{m} \\
&= \frac{2}{m} \cdot \text{ICost}_{\mathcal{D}_{\text{MC}}} (\pi_{\text{MC}}) + \frac{4}{m} = \frac{O(1)}{m} \cdot \text{ICost}_{\mathcal{D}_{\text{MC}}} (\pi_{\text{MC}})
\end{aligned}$$

where in the last inequality we used the fact that information cost of π_{MC} is at least 1. \blacksquare

We now have,

Theorem 5. *There exists a sufficiently small constant $\delta > 0$, such that for any $\omega(1/\sqrt{n}) \leq \varepsilon \leq o(1/\log n)$, and $m = \text{poly}(n)$,*

$$\text{CC}_{\mathcal{D}_{\text{MC}}}^{\delta}(\text{MaxCover}) = \Omega(m/\varepsilon^2).$$

Proof. Suppose there exists a δ -error protocol π_{MC} for MaxCover on \mathcal{D}_{MC} for a sufficiently small constant δ (to be determined later), with $\|\pi_{\text{GHD}}\| = o(m/\varepsilon^2)$; by Proposition 2.4, $\text{ICost}_{\mathcal{D}_{\text{MC}}}(\pi_{\text{MC}}) = o(m/\varepsilon^2)$ as well. Hence, by Lemma 4.5, we obtain a $(\delta + o(1))$ -error protocol π_{GHD} for GHD_{t_1} on \mathcal{D}_{GHD} with $\|\pi_{\text{GHD}}\| = o(m/\varepsilon^2)$ and $\text{ICost}_{\mathcal{D}_{\text{GHD}}^{\text{N}}}(\pi_{\text{GHD}}) = o(1/\varepsilon^2) = o(t_1)$. However, since $\|\pi_{\text{GHD}}\| = o(m/\varepsilon^2) = 2^{o(t_1)}$ as $m = \text{poly}(n)$ and $t_1 = \omega(\log n)$, we can now apply Lemma 4.2 and argue that $\text{ICost}_{\mathcal{D}_{\text{GHD}}^{\text{N}}}(\pi_{\text{GHD}})$ is $\Omega(t_1)$ (by taking δ smaller than the bounds in the Lemma 4.2); a contradiction with the information cost of π_{GHD} obtained by Lemma 4.5. \blacksquare

We point out that to extend the results in Theorem 5 to $\varepsilon > 1/\log n$ case (i.e., the case not handled by Theorem 5), we can simply use an existing $\Omega(m)$ lower bound of [42] (Theorem 21) for this range of the parameter ε .

We can now prove Theorem 4 by using Theorem 5, the same exact way as we proved Theorem 1, i.e., by defining a random partitioning version of the distribution \mathcal{D}_{MC} and proving the lower bound using that partitioning. We briefly sketch the proof here.

Proof Sketch of Theorem 4. Define the distribution \mathcal{D}'_{MC} similar to the distribution \mathcal{D}_{MC} with the difference that after creating the sets \mathcal{S} and \mathcal{T} , we randomly partition the sets between the players (i.e., assign each set to Alice w.p. 1/2 and o.w. to Bob). The same exact argument in Lemma 3.7, combined with Theorem 5 (instead of Theorem 3 in Lemma 3.7) now proves that for some sufficiently small constant $\delta > 0$, $\text{CC}_{\mathcal{D}'_{\text{MC}}}^{\delta}(\text{MaxCover}) = \Omega(m/\varepsilon^2)$.

Furthermore, any p -pass s -space streaming algorithm for maximum coverage on random arrival streams can be turned into an $O(s \cdot p)$ -bit communication protocol for MaxCover on \mathcal{D}'_{MC} (with the same error probability); see the proof of Theorem 1 for more details. This, together with the lower bound on the distribution \mathcal{D}'_{MC} implies that $s = \Omega(m/(\varepsilon^2 \cdot p))$ as desired. ■

Acknowledgements

I am grateful to my advisor Sanjeev Khanna for valuable discussions, and to Ehsan Emamjomeh-Zadeh and Sanjeev Khanna for carefully reading the paper and many helpful comments. I also thank the anonymous reviewers of PODS 2017 for many insightful comments and suggestions.

References

- [1] ALON, N., MATIAS, Y., AND SZEGEDY, M. The space complexity of approximating the frequency moments. In *STOC* (1996), ACM, pp. 20–29.
- [2] ANAGNOSTOPOULOS, A., BECCHETTI, L., BORDINO, I., LEONARDI, S., MELE, I., AND SANKOWSKI, P. Stochastic query covering for fast approximate document retrieval. *ACM Trans. Inf. Syst.* 33, 3 (2015), 11:1–11:35.
- [3] ASSADI, S., KHANNA, S., AND LI, Y. Tight bounds for single-pass streaming complexity of the set cover problem. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016* (2016), pp. 698–711.
- [4] AUSIELLO, G., BORRÀ, N., GIANNAKOS, A., LUCARELLI, G., AND PASCHOS, V. T. Online maximum k-coverage. *Discrete Applied Mathematics* 160, 13-14 (2012), 1901–1913.
- [5] BADANIDIYURU, A., MIRZASOLEIMAN, B., KARBASI, A., AND KRAUSE, A. Streaming sub-modular maximization: massive data summarization on the fly. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014* (2014), pp. 671–680.
- [6] BAR-YOSSEF, Z., JAYRAM, T. S., KUMAR, R., AND SIVAKUMAR, D. An information statistics approach to data stream and communication complexity. In *43rd Symposium on Foundations of Computer Science (FOCS 2002), 16-19 November 2002, Vancouver, BC, Canada, Proceedings* (2002), pp. 209–218.
- [7] BAR-YOSSEF, Z., JAYRAM, T. S., KUMAR, R., AND SIVAKUMAR, D. Information theory methods in communication complexity. In *Proceedings of the 17th Annual IEEE Conference on Computational Complexity, Montréal, Québec, Canada, May 21-24, 2002* (2002), pp. 93–102.

- [8] BARAK, B., BRAVERMAN, M., CHEN, X., AND RAO, A. How to compress interactive communication. In *Proceedings of the 42nd ACM Symposium on Theory of Computing, STOC 2010, Cambridge, Massachusetts, USA, 5-8 June 2010* (2010), pp. 67–76.
- [9] BATENI, M., ESFANDIARI, H., AND MIRROKNI, V. S. Almost optimal streaming algorithms for coverage problems. *CoRR abs/1610.08096* (2016).
- [10] BRAVERMAN, M. Interactive information complexity. In *Proceedings of the 44th Symposium on Theory of Computing Conference, STOC 2012, New York, NY, USA, May 19 - 22, 2012* (2012), pp. 505–524.
- [11] BRAVERMAN, M., GARG, A., PANKRATOV, D., AND WEINSTEIN, O. From information to exact communication. In *Symposium on Theory of Computing Conference, STOC’13, Palo Alto, CA, USA, June 1-4, 2013* (2013), pp. 151–160.
- [12] BRAVERMAN, M., GARG, A., PANKRATOV, D., AND WEINSTEIN, O. Information lower bounds via self-reducibility. In *Computer Science - Theory and Applications - 8th International Computer Science Symposium in Russia, CSR 2013, Ekaterinburg, Russia, June 25-29, 2013. Proceedings* (2013), pp. 183–194.
- [13] BRAVERMAN, M., AND RAO, A. Information equals amortized communication. In *IEEE 52nd Annual Symposium on Foundations of Computer Science, FOCS 2011, Palm Springs, CA, USA, October 22-25, 2011* (2011), pp. 748–757.
- [14] BRAVERMAN, M., AND WEINSTEIN, O. An interactive information odometer and applications. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015* (2015), pp. 341–350.
- [15] CHAKRABARTI, A., CORMODE, G., AND MCGREGOR, A. Robust lower bounds for communication and stream computation. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing, Victoria, British Columbia, Canada, May 17-20, 2008* (2008), pp. 641–650.
- [16] CHAKRABARTI, A., AND REGEV, O. An optimal lower bound on the communication complexity of gap-hamming-distance. In *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011* (2011), pp. 51–60.
- [17] CHAKRABARTI, A., SHI, Y., WIRTH, A., AND YAO, A. C. Informational complexity and the direct sum problem for simultaneous message complexity. In *42nd Annual Symposium on Foundations of Computer Science, FOCS 2001, 14-17 October 2001, Las Vegas, Nevada, USA* (2001), pp. 270–278.
- [18] CHAKRABARTI, A., AND WIRTH, A. Incidence geometries and the pass complexity of semi-streaming set cover. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016* (2016), pp. 1365–1373.
- [19] CHEN, J., NGUYEN, H. L., AND ZHANG, Q. Submodular maximization over sliding windows. *CoRR abs/1611.00129* (2016).
- [20] CHIERICHETTI, F., KUMAR, R., AND TOMKINS, A. Max-cover in map-reduce. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010* (2010), pp. 231–240.

- [21] CORMODE, G., KARLOFF, H. J., AND WIRTH, A. Set cover algorithms for very large datasets. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010* (2010), pp. 479–488.
- [22] COVER, T. M., AND THOMAS, J. A. *Elements of information theory* (2. ed.). Wiley, 2006.
- [23] DEMAINE, E. D., INDYK, P., MAHABADI, S., AND VAKILIAN, A. On streaming and communication complexity of the set cover problem. In *Distributed Computing - 28th International Symposium, DISC 2014, Austin, TX, USA, October 12-15, 2014. Proceedings* (2014), pp. 484–498.
- [24] DINUR, I., AND STEURER, D. Analytical approach to parallel repetition. In *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014* (2014), pp. 624–633.
- [25] DUBHASHI, D. P., AND PANCONESI, A. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, 2009.
- [26] EMEK, Y., AND ROSÉN, A. Semi-streaming set cover - (extended abstract). In *Automata, Languages, and Programming - 41st International Colloquium, ICALP 2014, Copenhagen, Denmark, July 8-11, 2014, Proceedings, Part I* (2014), pp. 453–464.
- [27] EPASTO, A., LATTANZI, S., VASSILVITSKII, S., AND ZADIMOGHADDAM, M. Submodular optimization over sliding windows. *CoRR abs/1610.09984, To appear in WWW (2017)* (2016).
- [28] FEIGE, U. A threshold of $\ln n$ for approximating set cover. *J. ACM* 45, 4 (1998), 634–652.
- [29] GÖÖS, M., JAYRAM, T. S., PITASSI, T., AND WATSON, T. Randomized communication vs. partition number. *Electronic Colloquium on Computational Complexity (ECCC) 22* (2015), 169.
- [30] GROSSMAN, T., AND WOOL, A. Computational experience with approximation algorithms for the set covering problem. *European Journal of Operational Research* 101, 1 (1997), 81–92.
- [31] GUHA, S., AND MCGREGOR, A. Stream order and order statistics: Quantile estimation in random-order streams. *SIAM J. Comput.* 38, 5 (2009), 2044–2059.
- [32] HAR-PELED, S., INDYK, P., MAHABADI, S., AND VAKILIAN, A. Towards tight bounds for the streaming set cover problem. In *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2016, San Francisco, CA, USA, June 26 - July 01, 2016* (2016), pp. 371–383.
- [33] IMPAGLIAZZO, R., AND KABANETS, V. Constructive proofs of concentration bounds. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, 13th International Workshop, APPROX 2010, and 14th International Workshop, RANDOM 2010, Barcelona, Spain, September 1-3, 2010. Proceedings* (2010), pp. 617–631.
- [34] INDYK, P., MAHABADI, S., AND VAKILIAN, A. Towards tight bounds for the streaming set cover problem. *CoRR abs/1509.00118* (2015).
- [35] INDYK, P., AND WOODRUFF, D. P. Tight lower bounds for the distinct elements problem. In *44th Symposium on Foundations of Computer Science (FOCS 2003), 11-14 October 2003, Cambridge, MA, USA, Proceedings* (2003), pp. 283–288.

- [36] JOHNSON, D. S. Approximation algorithms for combinatorial problems. *J. Comput. Syst. Sci.* 9, 3 (1974), 256–278.
- [37] KAPRALOV, M., KHANNA, S., AND SUDAN, M. Approximating matching size from random streams. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2014, Portland, Oregon, USA, January 5-7, 2014* (2014), pp. 734–751.
- [38] KARP, R. M. Reducibility among combinatorial problems. In *Proceedings of a symposium on the Complexity of Computer Computations, held March 20-22, 1972, at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York.* (1972), pp. 85–103.
- [39] KONRAD, C., MAGNIEZ, F., AND MATHIEU, C. Maximum matching in semi-streaming with few passes. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques - 15th International Workshop, APPROX 2012, and 16th International Workshop, RANDOM 2012, Cambridge, MA, USA, August 15-17, 2012. Proceedings* (2012), pp. 231–242.
- [40] KUSHILEVITZ, E., AND NISAN, N. *Communication complexity*. Cambridge University Press, 1997.
- [41] LUND, C., AND YANNAKAKIS, M. On the hardness of approximating minimization problems. *J. ACM* 41, 5 (1994), 960–981.
- [42] MCGREGOR, A., AND VU, H. T. Better streaming algorithms for the maximum coverage problem. *CoRR abs/1610.06199. To appear in ICDT (2017)* (2016).
- [43] MOSHKOVITZ, D. The projection games conjecture and the np-hardness of $\ln n$ -approximating set-cover. *Theory of Computing* 11 (2015), 221–235.
- [44] MUTHUKRISHNAN, S. Data streams: Algorithms and applications. *Foundations and Trends in Theoretical Computer Science* 1, 2 (2005).
- [45] NISAN, N. The communication complexity of approximate set packing and covering. In *Automata, Languages and Programming, 29th International Colloquium, ICALP 2002, Malaga, Spain, July 8-13, 2002, Proceedings* (2002), pp. 868–875.
- [46] PANCONESI, A., AND SRINIVASAN, A. Randomized distributed edge coloring via an extension of the chernoff-hoeffding bounds. *SIAM J. Comput.* 26, 2 (1997), 350–368.
- [47] SAHA, B., AND GETOOR, L. On maximum coverage in the streaming model & application to multi-topic blog-watch. In *Proceedings of the SIAM International Conference on Data Mining, SDM 2009, April 30 - May 2, 2009, Sparks, Nevada, USA* (2009), pp. 697–708.
- [48] SLAVÍK, P. A tight analysis of the greedy algorithm for set cover. *J. Algorithms* 25, 2 (1997), 237–254.
- [49] WEINSTEIN, O., AND WOODRUFF, D. P. The simultaneous communication of disjointness with applications to data streams. In *Automata, Languages, and Programming - 42nd International Colloquium, ICALP 2015, Kyoto, Japan, July 6-10, 2015, Proceedings, Part I* (2015), pp. 1082–1093.
- [50] YAO, A. C. Some complexity questions related to distributive computing (preliminary report). In *Proceedings of the 11th Annual ACM Symposium on Theory of Computing, April 30 - May 2, 1979, Atlanta, Georgia, USA* (1979), pp. 209–213.

A Tools from Information Theory

Here, we briefly introduce some basic facts from information theory that are needed in this paper. We refer the interested reader to the textbook by Cover and Thomas [22] for an excellent introduction to this field.

We use the following basic properties of entropy and mutual information (proofs can be found in [22], Chapter 2).

Fact A.1. *Let A , B , and C be three (possibly correlated) random variables.*

1. $0 \leq \mathbb{H}(A) \leq |A|$. $\mathbb{H}(A) = |A|$ iff A is uniformly distributed over its support.
2. $\mathbb{I}(A : B) \geq 0$. The equality holds iff A and B are independent.
3. Conditioning on a random variable reduces entropy: $\mathbb{H}(A | B, C) \leq \mathbb{H}(A | B)$. The equality holds iff $A \perp C | B$.
4. The chain rule for mutual information: $\mathbb{I}(A, B : C) = \mathbb{I}(A : C) + \mathbb{I}(B : C | A)$.

We also use the following two simple facts, which assert conditions in which conditioning can provably increase (resp. decrease) the mutual information.

Fact A.2. *For random variables A, B, C, D , if $A \perp D | C$, then $\mathbb{I}(A : B | C) \leq \mathbb{I}(A : B | C, D)$.*

Proof. Since A and D are independent conditioned on C , by Fact A.1-(3), $\mathbb{H}(A | C) = \mathbb{H}(A | C, D)$ and $\mathbb{H}(A | C, B) \geq \mathbb{H}(A | C, B, D)$. We have,

$$\begin{aligned} \mathbb{I}(A : B | C) &= \mathbb{H}(A | C) - \mathbb{H}(A | C, B) = \mathbb{H}(A | C, D) - \mathbb{H}(A | C, B) \\ &\leq \mathbb{H}(A | C, D) - \mathbb{H}(A | C, B, D) = \mathbb{I}(A : B | C, D) \end{aligned}$$

■

Fact A.3. *For random variables A, B, C, D , if $A \perp D | B, C$, then, $\mathbb{I}(A : B | C) \geq \mathbb{I}(A : B | C, D)$.*

Proof. Since $A \perp D | B, C$, by Fact A.1-(3), $\mathbb{H}(A | B, C) = \mathbb{H}(A | B, C, D)$. Moreover, since conditioning can only reduce the entropy (again by Fact A.1-(3)),

$$\begin{aligned} \mathbb{I}(A : B | C) &= \mathbb{H}(A | C) - \mathbb{H}(A | B, C) \geq \mathbb{H}(A | D, C) - \mathbb{H}(A | B, C) \\ &= \mathbb{H}(A | D, C) - \mathbb{H}(A | B, C, D) = \mathbb{I}(A : B | C, B) \end{aligned}$$

■

Finally, we use the following simple inequality that states that conditioning on a random variable can only increase the mutual information by the entropy of the conditioned variable.

Fact A.4. *For any random variables A, B and C , $\mathbb{I}(A : B | C) \leq \mathbb{I}(A : B) + \mathbb{H}(C)$.*

Proof.

$$\begin{aligned} \mathbb{I}(A : B | C) &= \mathbb{I}(A : B, C) - \mathbb{I}(A : C) \\ &= \mathbb{I}(A : B) + \mathbb{I}(A : C | B) - \mathbb{I}(A : C) \\ &\leq \mathbb{I}(A : B) + \mathbb{H}(C | B) \leq \mathbb{I}(A : B) + \mathbb{H}(C) \end{aligned}$$

where the first two equalities are by chain rule (Fact A.1-(4)), the second inequality is by definition of mutual information and its positivity (Fact A.1-(2)), and the last one is because conditioning can only reduce the entropy (Fact A.1-(3)). ■

B Proof of Lemma 4.2

Proof. The proof consists of two separate parts. We first prove that there exists a pair $a, b \in [t]$, for which GHD is still “hard” under the distribution $\mathcal{U}(a, b) := \mathcal{U} \mid |A| = a, |B| = b$ (i.e., when we fix the size of the sets A and B), and in next part, use this fact to prove the bound for the distribution $\mathcal{D}_{\text{GHD}}^N$ defined for the same pair of a, b found in the first part (the proof of second part is basically the same as Lemma 3.5).

Claim B.1. *Let $\delta > 0$ be a sufficiently small constant; there exists a pair of $a, b \in [t]$ such that*

$$\text{IC}_{\mathcal{U}(a,b)}^\delta(\text{GHD}) = \Omega(t)$$

whereby $\mathcal{U}(a, b) = \mathcal{U} \mid |A| = a, |B| = b$.

Proof. Let δ be as in Lemma 4.1. Suppose by contradiction that for all $a, b \in [t]$, $\text{IC}_{\mathcal{U}(a,b)}^\delta(\text{GHD}) = o(t)$, and let $\pi_{a,b}$ be the protocol achieving this bound for a specific choice of a, b . We design the following protocol π for GHD on the distribution \mathcal{U} : Given an input $(A, B) \sim \mathcal{U}$, Alice and Bob first communicate $a = |A|$, $b = |B|$ to each other and then run $\pi_{a,b}$ on their input and output the same answer as in $\pi_{a,b}$. Since each $\pi_{a,b}$ is computed on the same exact distribution as $\mathcal{U}(a, b)$ (corresponding to the same parameters a and b), π is a δ -error protocol for GHD on \mathcal{U} . We now bound the information cost of π as follows (in the following, Π corresponds to the protocol π , $\Pi_{a,b}$ corresponds to the protocol $\pi_{a,b}$, and X_A (resp. X_B) is a random variable for size of A (resp. B))

$$\begin{aligned} \text{ICost}_{\mathcal{U}}(\pi) &= \mathbb{I}(A : \Pi \mid B) + \mathbb{I}(B : \Pi \mid A) \\ &= \mathbb{I}(A : \Pi_{a,b}, X_A, X_B \mid B) + \mathbb{I}(B : \Pi_{a,b}, X_A, X_B \mid A) \\ &= \mathbb{I}(A : \Pi_{a,b} \mid B, X_A, X_B) + \mathbb{I}(B : \Pi_{a,b} \mid A, X_A, X_B) + 2 \cdot \mathbb{H}(X_A, X_B) \\ &\quad \text{(by Fact A.1-(4)) and Fact A.1-(2)} \\ &= \mathbb{E}_{(a,b)} \left[\mathbb{I}(A : \Pi_{a,b} \mid B, X_A = a, X_B = b) + \mathbb{I}(B : \Pi_{a,b} \mid A, X_A = a, X_B = b) \right] + O(\log t) \\ &\quad \text{(by Fact A.1-(1))} \\ &= \mathbb{E}_{(a,b)} \left[\mathbb{I}_{\mathcal{U}(a,b)}(A : \Pi_{a,b} \mid B) + \mathbb{I}_{\mathcal{U}(a,b)}(B : \Pi_{a,b} \mid A) \right] + O(\log t) \\ &\quad \text{(by definition, } \mathcal{U}(a, b) := \mathcal{U} \mid X_A = a, X_B = b) \\ &= \mathbb{E}_{(a,b)} \left[\text{ICost}_{\mathcal{U}(a,b)}(\pi_{a,b}) \right] + O(\log t) \quad \text{(by definition of information cost of } \pi_{a,b}) \\ &= o(t) + O(\log t) = o(t) \end{aligned}$$

where in the second last inequality we used assumption that $\text{ICost}_{\mathcal{U}(a,b)}(\pi_{a,b}) = o(t)$ for all a and b .

Consequently, we obtained a δ -error protocol π for GHD on the distribution \mathcal{U} with information cost of $o(t)$, a contradiction with Lemma 4.1. This means that there should exist at least one pair a, b such that $\text{IC}_{\mathcal{U}(a,b)}^\delta(\text{GHD})$ is $\Omega(t)$, proving the claim. \blacksquare

Now fix a and b as in Claim B.1 and define \mathcal{D}_{GHD} accordingly. Suppose by contradiction that $\text{ICost}_{\mathcal{D}_{\text{GHD}}}(\pi_{\text{GHD}})$ is some $\tau = o(t)$. We can use the previous information odometer argument (i.e., Lemma 3.6) to create a protocol π'_{GHD} that solves GHD on the distribution $\mathcal{U}(a, b)$ and has information cost of $\text{ICost}_{\mathcal{U}(a,b)}(\pi'_{\text{GHD}}) = o(t)$, a contradiction with Claim B.1. We can simply create π'_{GHD} as follows: run the protocol π_{GHD} and the information odometer in parallel; whenever the information cost of π_{GHD} is larger than $c \cdot \tau$ (for a sufficiently large constant c), terminate the protocol and output an arbitrary answer, otherwise output the same answer as π_{GHD} . This ensures that $\text{ICost}_{\mathcal{U}(a,b)}(\pi'_{\text{GHD}}) = o(t)$. By definition of the distribution \mathcal{D}_{GHD} , and the fact that

$\text{ICost}_{\mathcal{D}_{\text{GHD}}}(\pi_{\text{GHD}}) = \tau$, in the cases that we terminate the protocol π_{GHD} , the answer to **GHD** can be arbitrary w.p. $1 - o(1)$ and hence the new protocol is $(\delta + o(1))$ -error protocol for **GHD** on $\mathcal{U}(a, b)$. This can be made formal exactly as in the proof of Lemma 3.6. The rest of the proof now follows from Lemma 3.6 exactly as in Lemma 3.5. ■